Process analysis for marketing research

Pieters, Constant

DOI:
10.26116/center-lis-2009

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Process Analysis for Marketing Research

CONSTANT PIETERS

Process Analysis for Marketing Research

Proefschrift ter verkrijging van de graad van doctor aan Tilburg University

op gezag van de rector magnificus, prof. dr. W.B.H.J. van de Donk, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de Aula van de Universiteit op maandag 14 december 2020 om 13.30 uur

door

Constant Pieters

geboren te Oostburg, Nederland.

Promotor:

Prof. dr. F.G.M. Pieters, Tilburg University

Copromotor:

Dr. A. Lemmens, Erasmus University Rotterdam

Promotiecommissie:

Prof. dr. H. Baumgartner, Pennsylvania State University

Prof. dr. T.H.A. Bijmolt, Rijksuniversiteit Groningen

Prof. dr. B. Deleersnyder, Tilburg University

Prof. dr. I. Geyskens, Tilburg University

Prof. dr. E. Gijsbrechts, Tilburg University

Prof. dr. ir. P.W.J. Verlegh, Vrije Universiteit Amsterdam

Preface

It still feels quite surrealistic to present the final version of this dissertation, which marks the end of the Ph.D. track that I started in 2014. This dissertation is a report of my doctoral research and consists of three essays that apply, compare, and attempt to extend process analysis methodologies for marketing research. I hope that you will enjoy reading about process theories and models, statistics, and even guinea pigs. Below, I would like to take the opportunity to reflect on my Ph.D. journey and thank those who made it possible.

First and foremost, I express my deepest gratitude to my advisors Aurélie and Rik. Throughout the years, you both contributed in uncountable and complementary ways to my professional and personal development. I absolutely enjoy working with both of you. Thank you for your feedback, advice, and continuous support.

Aurélie, I vividly remember our first meeting in 2013. I was quite nervous but you were friendly and relaxing, as you always are. I really appreciate your prevailing positivity and encouragement, especially at nerve-racking moments just before presentations or submissions. Thank you so much for getting me on board the Ph.D. program, your commitment to our research, and your incredible generosity in terms of time and other resources.

Rik, you often mention that although we are not related in terms of family, we are in spirit. I completely agree with you. Your dedication to your work is contagious, and I have learned so much from our research and Team Pieters teaching. I truly appreciate the directness of your feedback, your commitment to my intellectual and personal development (are we at version 4.0 already?), and your excellent referrals of hoppy, fermented barley beverages.

Besides amazing mentors, I was honored to have a distinguished Ph.D. committee. I am grateful for your effort, your comments, and support. Barbara, thank you for your

feedback and coaching as coordinator of the program. Els, you master the art of raising critical issues in a very nice and constructive way, thank you for your feedback throughout the years. Hans, your expertise on process analysis is truly inspiring, thanks a lot for sharing your insights. Inge, thank you for your support during the Research MSc., your enthusiasm for my teaching, and your help when I was on the job market. Peeter, thank you for the encouragement and for stimulating me to dig deeper in referral theory. Tammo, thank you for your comments and inviting me to contribute to your EMAC special session in 2018.

It was a pleasure to study and work at the Tilburg Department of Marketing. I really appreciate the opportunities I was given, the stimulating environment, and the teamwork. Many thanks to all colleagues for the lectures, feedback during the summer camps, and the numerous chats. I am sad to leave. To my fellow Ph.D. candidates, thank you for sharing offices and the ups and downs of the program with me. Special thanks go to the team of marketing lecturers for welcoming me in 2018, your dedication to teaching is inspirational. Thank you Aniek, Elke, Hendrik and Teun for co-teaching, it was a pleasure and I have learned a lot from you. Aukje, Carlie, and Giuli, I am grateful for your help. Thank you Heidi, Nancy, Scarlett, for the support. I thank the CentER graduate officers for their administrative support. Stereotypically, it is rare for a graduate student to turn down free food. Thank you Ana, Francesca, Lucas, and others for the chocolate. It really helped.

Several others deserve a special mention. Thank you Ernst for your support to enroll in the Research MSc. Many thanks go to Richard, thank you for the countless chats and for surviving grad school together. It was amazing to work with you, and I hope we can write another paper in the future. Thank you Zi-Lin for your mentorship, I am super proud of our work. I would like to thank Jacob Goldenberg and lab for hosting me at IDC Herzliya in the fall of 2015. Anatoli, Andreas, Danny, Jonne and Radek, thanks for the many stimulating chats and drinks over the years. Thank you Maxime for collaborating. Thank you Stefano

Puntoni for sharing your perspectives on the job market. I am grateful to Jack, John, Harald, Hauke, Ting and Valentyna for hosting me at UNSW in September 2019.

Although it is often quite confronting to Ph.D. candidates to be asked about their progress, my family, in-laws and friends consistently provided me a tremendous amount of social support. Thank you very much. I wish I had more space to properly express my gratitude to each of you. I owe a lot to my parents Paul and Suzy and I am really grateful for their endless support of me and my education. Thank you Jan, José, Lotte and Rick for welcoming me into the family and helping me to relax and unwind. Thank you Joep for including me in the group, thank you all for the friendship and many chats. Lara, thank you for your friendship throughout the years. Thank you Alen, Mark, Peter, Thomas, Tom and Tom for keeping me sane.

Fenna, you are usually more confident in me than I am in myself. You were even prepared to marry me and move across the world together. I am truly convinced we complement each other in numerous ways, and I cannot thank you enough for everything you do. You are my best friend and I love you.

CONSTANT PIETERS
Tilburg, The Netherlands
August 2020

Contents

Chapter 1 – General Introduction

1.1 Guinea Pigs

Suppose that an analyst is interested in estimating the relative impact of hereditary, environmental, and other factors on the transmission of fur color between generations of guinea pigs or their birth weight. Rightfully so, because studying the relative importance of the effects that inputs have on relevant outcomes is one of the main objectives of scientific inquiry. You might wonder why the opening example of this introduction is about guinea pigs. Indeed, an investigation towards the determinants of guinea pig fur color and birth weight seems distant from conventional topics in marketing research. Yet, guinea pigs quite literally stood at the inception of process analysis methodologies that are currently widespread in the marketing discipline and the social sciences more generally. Substantive questions about the genetics of guinea pigs stimulated Sewall Wright (1889-1988), an American geneticist, to make important contributions to process analysis methodologies during and after his years as a graduate student at Harvard.

In 1914, Wright was assigned by his Ph.D. advisor William Castle to use Karl Pearson's partial correlation coefficient to reanalyze five bone length measures of rabbits and decompose the variation in measurements into general and specific size factors of the animals (Provine 1989, pp. 78-79). Castle was impressed by the correlation analysis, which ultimately led to an "…attempt to assign definite values to the different classes of growth factors which are indicated" (Wright 1918, p. 370). Wright's early work already hinted at the distinction between input and output variables, and partitioned variance similar to factor analysis (Bollen 1989, p. 5), yet it did not formally propose process analysis. Wright wanted more (Provine 1989, pp. 79 & 127-128), and he continued to explore how to quantify the relative effects of inputs on outputs. This resulted in the development of the path coefficient, one of Wright's most important contributions to process analysis.

When Wright was a master's student at the University of Illinois and met his Ph.D. advisor, it was made clear that he would inherit a colony of guinea pigs when Castle's assistant and graduate student John Detlefsen left (Provine 1989, p. 80). Wright continued tending for the colony during his years at the United States Department of Agriculture (USDA), until his retirement at the University of Chicago in 1954, remaining on the faculty of the University of Wisconsin until 1960 (Crow 1992). He used data from the colony to present an analysis that aimed to quantify the impact of hereditary, environmental and other factors on the transmission of fur color between generations of guinea pigs (Wright 1920).

That 1920 article presented all elements of contemporary path analysis (Bollen 1989; Wolfle 1999). First, it formally introduced the path coefficient, which was characterized as the sum of the paths that connected two variables. It quantified the relative importance of the effects of inputs on outputs. Second, the product of the path coefficients constituted the contribution of inputs with effects through intervening. The most important result was, in Wright's own words, that "[t]he correlation between two variables can be shown to equal the sum of the products of the chains of path coefficients along all of the paths by which they are connected" (Wright 1920, p. 330). Third, the article presented graphical diagrams that clearly identified how inputs, throughputs and outputs are expected to be related (p. 328). It concluded that variations in fur color of guinea pigs were determined for about 3% by heredity in an inbred stock of guinea pigs, but for 42% in a control stock (Wright 1920). A follow-up article concluded that the effect of the size of litter on the weight of guinea pigs at birth and at weaning (33 days) was found to be larger though a reduced fetus growth rate than through its influence on early birth (Wright 1921).

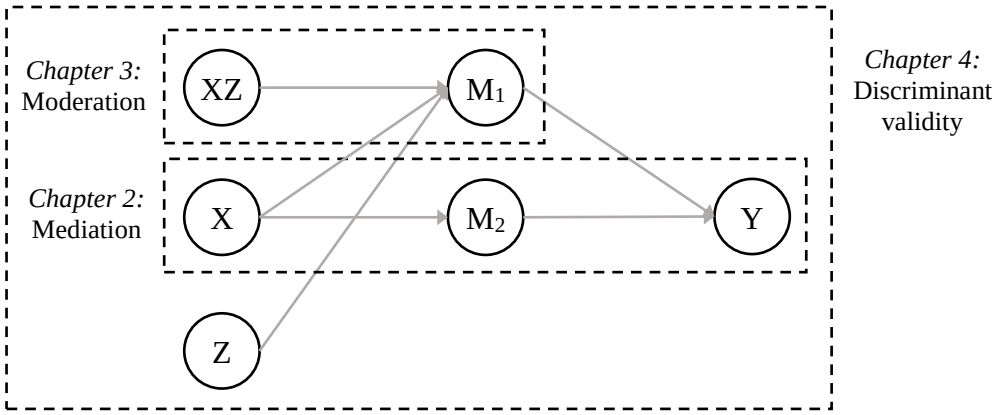
Initially, Wright's colleagues at the USDA were not enthusiastic about his novel methods and findings (Provine 1989, p. 134). Moreover, the ideas were criticized by Niles (1922), and endured an intense controversy between Wright and Sir Ronald Fisher (Provine

1992). Interestingly, early applications in the social sciences can be traced back to Burks (1928), who concluded that 33% of the variance in a child's intelligence could be explained by hereditary factors, and 4% by environmental factors. Yet, it took over 40 years for Wright's contributions to be (re)discovered and popularized in sociology (Duncan 1966), which led to further dissemination in the social sciences.

1.2 Process Theories and Analyses

Fast forward to 2020, which marks the centennial anniversary of Wright's (1920) contributions, process analysis has become an indispensable tool to provide insights in the relative contributions of the effects that inputs have on outputs, and process theories in general. Commonly, marketing researchers and managers are not only interested in to what extent input variables (X) have simple effects on outcomes (Y). Instead, they are often interested in quantifying *how* and *when* input variables affect outcomes (Spencer et al. 2005). This dissertation defines a *process theory* as a theory that aims to establish how and/or when one or more input variables influence one or more outcomes. *Process models* depict these

Figure 1.1
Hypothetical Process Model with the Focus of the Remaining Chapters



Notes: Circles refer to constructs, their indicators are omitted for exposition. X and Z are inputs, Z is a moderator and XZ is the interaction between X and Z. Ms are mediators and Y is the outcome. Arrows are causal relationships between constructs, direct paths from inputs to M and Y and correlations between Ms are omitted for brevity. Dashed boxes with annotations indicate the parts of the hypothetical process model that Chapters 2 to 4 focus on.

theories in equations and graphical representations, such as those introduced by Wright (1920).

Process analysis takes process theories to data and aims to empirically identify and quantify the relative importance of the pathways that are presented by process theories and models. It commonly makes use of statistical methods such as analysis of variance (ANOVA), regression, path analysis, structural equation modeling (SEM), and so forth.

Figure 1.1 presents a visual representation of a hypothetical process model. Circles refer to constructs, and arrows are relationships between them. It is common to specify a mediator to answer the question: "How does X affect Y?" A mediator (M) is then a throughput variable of the X-Y relationship. Wright (1920) already accounted for mediation with his proposed method of multiplying path coefficients, later referring to "intervening" variables (p. 163) or relationships that could be affected by "mediation" (Wright 1934, p. 179). Little has changed to the core principle of mediation analysis: the indirect effect of one variable on another is captured by the product of the path weights connecting the two variables (Pieters 2017, p. 693).

Moderation answers the question: "When does X affect Y?" A moderator (Z) is a condition or contingency that strengthens or weakens the X-Y effect. Statistically, moderation refers to an interaction, here XZ is the multiplication between variable X and Z. Although Wright did not explicitly develop moderation, its importance was recognized by him in that "…one often has to deal with a group of characteristics or conditions which are correlated because of a complex of *interacting*, uncontrollable, and often obscure causes" (Wright 1921, p. 557, emphasis added). Saunders (1955), using the author's own words, "christened" (p. 54) the moderator as a useful tool for prediction, although it was earlier also referred to as a "population control variable" (Gaylord and Carroll 1948) or "joint causation" (Court 1930). Later, the moderator-mediator distinction was elaborated on in one of the most

cited articles in psychology to date (Baron and Kenny 1986; as of February 2020 cited 90,443 times according to Google Scholar, and 39,955 times according to Web of Science). A process model contains mediation, moderation, or a combination of mediation and moderation like in Figure 1.1. This combination can be referred to as conditional process analysis (Hayes and Preacher 2013).

Process theories and analyses are widespread in contemporary marketing research. For example, an editor of the *Journal of Consumer Research (JCR)* noted, anecdotally, that the majority of submitted manuscripts propose a new phenomenon and demonstrate the process by which it may occur by testing for mediation, moderation, and boundary conditions (Deighton et al. 2010). As an example of mediation, customer participation increases customer empowerment and customer satisfaction which in turn affect firm performance (Auh et al. 2019). Or, for another illustration, the effect of consumer busyness and lack of leisure time on perceived status is mediated by human capital characteristics and perceived scarcity (Bellezza et al. 2017). As an example of moderation, the effect of brand differentiation on profits is moderated by market uncertainty. When market uncertainty increases, the positive effect of brand differentiation on profits increases (Dahlquist and Griffith 2014). Similarly, the effect of brand extension fit on brand extension success depends on the quality of the parent brand, that is, the positive effect of the parent brand on extension success increases as the fit between parent brand and extension product increases (Völckner and Sattler 2006). Recently, Pieters (2017) found that 86 of the 121 articles (71%) that used experiments in volumes 41 and 42 of *JCR* (2014-2016) contained at least one mediation analysis. Out of the 166 mediation analyses investigated, 82 (49%) examined a combination of moderation and mediation and 29 (17%) had multiple mediators.

Process theories with mediators and moderators have a large academic and practical relevance (Spencer et al. 2005). Establishing mediation provides evidence for the otherwise

hidden intervening mechanisms in theories. Moreover, insights in mediators through which marketing interventions lead to performance outcomes enables managers to better gauge the effectiveness of such interventions. It gives managers additional tools to intervene in the multiple paths that drive performance (e.g., X to M, M to Y as well as X to Y). The nuanced insights from process evidence facilitate interventions that would be hidden by a focus on the total effect of X on Y.

Moderation identifies the boundary conditions and generalizability of purported theories (Goldsby et al. 2013). For instance, if an effect is weaker for individuals with a certain trait or in a certain state, the implication is that processes related to the trait or state drive the effect (Kahn et al. 2006). Moreover, moderation provides managers insights in the conditions under which marketing interventions yield their largest effects. Insights in moderation effects aid firms in using the right treatment in the right situation or for the right customer segment. Moderation explains why interventions can at times fail to achieve the desired results but lead to favorable performance outcomes in other situations or for specific segments. In sum, insights in the processes contribute to richer theories and more effective marketing interventions.

1.3 Process Analysis for Marketing Research: Roadmap

This dissertation contains three essays (Chapters 2 to 4) on process analysis for marketing research and the final Chapter 5 summarizes, has follow-up analyses, and concludes. Figure 1.1 visualizes the components of the hypothetical process model that the chapters have a primary focus on.

Chapter 2 applies mediation methods to a substantive question. It examines consumer referrals and focuses on the referral reinforcement effect: referred customers have a higher inclination of making referrals than non-referred customers have. Four studies (an analysis of ridesharing customers, a reanalysis of published data from a bank's referral program, a new

survey among moviegoers, and a controlled experiment using a Super Bowl commercial) quantify the referral reinforcement effect across contexts (organic vs. incentivized referrals), using different methodologies. Mediation analyses decompose the referral reinforcement effect into satisfaction-mediated and non-satisfaction-mediated parts. A final study explores customer lay beliefs about potential drivers of the referral reinforcement effect.

Chapter 3 compares existing moderation methods in the face of random measurement error, which is common in marketing research. It focuses on six methods that differ in how measurement error is accounted for. The chapter reviews the usage of these methods in marketing research. Two of the methods, means and multi-group, are widely used but do not account for measurement error. The other methods, including factor scores, corrected means, product indicators, and latent product, account for measurement error but have hardly been used so far. The disproportionate use of the means and multi-group methods calls for an assessment of the performance of these approaches relative to theoretically superior approaches. Monte Carlo simulations quantify the bias and statistical power of the estimated moderation effect for each of the six methods, using the results from the literature review as input. The chapter concludes with recommendations for usage of the methods.

Chapter 4 is an attempt to extend existing discriminant validity methods that examine whether measures of theoretically distinct constructs are empirically distinct. Measure distinctiveness is a necessary condition to establish construct validity and thus for meaningful theory-testing. Yet, process analyses can be at risk for not meeting discriminant validity. For example, sequential mediators are by definition hypothesized to be strongly related and mediators in parallel might correlate highly if they capture fine-grained processes that cannot be empirically distinguished. Unfortunately, discriminant validity is rarely assessed in marketing research. Even more, discussions of discriminant validity to date have exclusively focused on bivariate discriminant validity, which captures the empirical distinctiveness

within each pair of measures of constructs. Chapter 4 provides a framework of discriminant validity and a new multivariate discriminant validity criterion. The multivariate criterion accounts for all correlations between measures of constructs in a set instead of assessing pairs of measures. Chapter 4 explores sets of up to four measures of constructs. Then, it provides a quantitative literature review and meta-analysis of multiple mediation process models in marketing, to illustrate discriminant validity assessment in an important theory testing domain. Four case studies demonstrate situations that are of particular risk of lack of discriminant validity. They cast doubt on the validity of the purported multiple mediation theories. An online application is developed to increase the accessibility of the discriminant validity criteria.

Chapter 5 provides a general discussion that first gives an overview of the results of Chapters 2 to 4. It then presents three follow-up studies that address remaining issues and it concludes by speculating about the road ahead for process analysis.

1.4 Overview of Themes

Overall, this dissertation presents three essays on process analysis and its preconditions. Table 1.1 demarcates and gives an overview of substantive themes that are discussed in each chapter. Mediation and moderation return throughout this dissertation. It explores applications of mediation (Chapters 2 and 4), and compares existing moderation methods (Chapter 3). Chapter 2 applies moderation. The concluding Chapter 5 follows up.

The chapters treat all facets of construct validity (Peter 1981), the evaluation of the extent to which a measure assesses the construct it is deemed to measure (Strauss and Smith 2009, p. 2). The dissertation discusses reliability, a first aspect of construct validity, throughout. When applicable, it is assumed that the observed variance in a measure (X) is equal to sum of the variance of the true score (T_X) and random and independent measurement error (ε_X), formally: $var(X) = var(T_X) + var(\varepsilon_X)$. The random measurement error is accounted

Table 1.1
Overview of the Themes in the Chapters in this Dissertation

	Discussed in Chapter			
Theme	2	3	4	5
Process analysis				
Mediation				
Application of existing methods	✓	×	✓	×
Comparison of existing methods	×	×	×	×
Extension of existing methods	×	×	×	×
Moderation				
Application of existing methods	✓	×	×	×
Comparison of existing methods	×	✓	×	✓
Extension of existing methods	×	×	×	×
Construct validity				
Reliability				
Application of existing methods	✓	✓	✓	✓
Comparison of existing methods	×	×	×	×
Extension of existing methods	×	×	×	×
Convergent validity				
Application of existing methods	✓	×	×	×
Comparison of existing methods	×	×	×	×
Extension of existing methods	×	×	×	×
Discriminant validity				
Application of existing methods	✓	×	✓	✓
Comparison of existing methods	×	×	✓	×
Extension of existing methods	×	×	✓	×
Nomological validity				
Application of existing methods	✓	×	×	×
Comparison of existing methods	×	×	×	×
Extension of existing methods	×	×	×	×
Data and measurement (model)				
Summary statistics data (SSD)	✓	✓	✓	✓
Multidimensional measurement	✓	×	✓	×
Single-indicator measurement	✓	✓	✓	✓
Systematic measurement error	✓	✓	✓	×
Non-normality in variables	×	×	×	✓
Structural model				
Multicollinearity	×	✓	✓	✓
Non-linear models (e.g., probit)	✓	×	×	×
U-shapes	✓	×	×	✓

Notes: Table contains themes and checkmarks to outline in which chapters of the dissertation the themes are discussed. The checkmark ✓ means that the theme is discussed, and × that it is not.

for throughout the chapters. Reliability is then (an estimate of) the proportion of true score variance in the observed measure. The discussion sections of Chapters 2 to 4 discuss systematic, non-random, measurement error for instance due to common method variance (CMV). Chapter 2 explores convergent validity by generalizing the referral reinforcement effect using different customer satisfaction measures. Chapter 2 assesses discriminant validity using established criteria, Chapter 4 is an attempt to extend these criteria, and Chapter 5 follows up by reassessing the evidence for discriminant validity in the data of Chapter 2 using the new criteria. Finally, Chapter 2 focuses on nomological validity by investigating the relationships between measures of constructs that are theoretically expected to be related.

Turning to the data, measurement, and the measurement model, all chapters use summary statistics data (SSD), which are a compact, aggregate, form of raw data that can readily be included in analysis reports (Pieters 2017). Chapters 2 and 4 treat multidimensional measurement of customer satisfaction and market-orientation respectively, the remaining chapters focus on unidimensional measurement. All chapters deal with single-indicator as well as multi-indicator measurement. Non-normality in latent variables is discussed in Chapter 5 in the context of the moderation methods presented in Chapter 3.

In the structural model, multicollinearity, correlation between explanatory variables, plays a role in the context of moderation methods (Chapters 3 and 5), discriminant validity (Chapter 4), and statistical power (Chapter 5). The presented structural models are linear, except for the probit models in Chapter 2. Yet, there is little reason to expect that the process analysis methodologies presented in the chapters do not generalize to non-linear models. Chapter 2 estimates a U-shaped relationship (Haans et al. 2016) as a robustness check. U-shapes return in Chapter 5 as generalizations of moderation analysis.

Chapter 2 – The Referral Reinforcement Effect: Being Referred Increases Customers' Inclination to Refer in Turn1

2.1 Introduction

When Tesla launched its 2019 referral reward program (RRP), it offered Tesla car owners 1,000 miles of free supercharging, plus a chance to win an exclusive Tesla car each time a friend used a referral code (Tesla 2019). Here, referrals – incentivized or not – are cast as explicit, positive, peer-to-peer "buy" advisories from existing customers to prospective ones – quite distinct from mere brand-related discussions, mere mentions, general reviews, and observational learning (Berger 2014). Referrals have become an essential source of growth for firms like Tesla, Dropbox, Airbnb and Uber. A webhosting company study of customer acquisition by Villanueva et al. (2008) reported weekly inflows of new customers acquired via referrals doubling those acquired by traditional marketing instruments. They opined this was due to a *reinforcement effect*: customers acquired by referrals being more prone to refer than customers acquired by other means. In case such a referral reinforcement effect is sizable and reliable across industries, many firms might be undervaluing referrals.

Recent work has found that referred customers tend to have higher customer lifetime values (CLV) (Schmitt et al. 2011; Van den Bulte et al. 2018). Adding referral reinforcement effects would further imply that referred customers also yield higher referral values, making them even more valuable as referral transmitters through social networks, thereby triggering referral cascades (Goel et al. 2015; Leskovec et al. 2007). The total profitability of RRPs could thus exceed prior estimates. In fact, the return on investment of a referral reward should logically take into account both customers directly acquired through referrals, as well as the stream of subsequent acquisitions due to the increased share of referrals in the customer base.

[1] Maxime C. Cohen (McGill University) provided access to data used in Cohen et al. (2019) for Study 1.

Despite their potential managerial importance, referral reinforcement effects have attracted surprisingly sparse theorizing and research. True, much is known about various drivers of the likelihood of making referrals such as customer satisfaction and loyalty (Anderson 1998; De Matos and Rossi 2008), opinion leadership (Iyengar et al. 2011), age and income (Kumar et al. 2010), self- and other-directed motives (Berger 2014; Engel et al. 1969), and monetary and other incentives (Ahrens et al. 2013; Jin and Huang 2014; Verlegh et al. 2013) as exemplified by the above Tesla case (see Kumar et al. (2010) for an extensive overview of drivers). Yet, the effect *per se* of referral reception on a customer's inclination to refer others is, to our knowledge, largely uncharted. In fact, studies documenting potential referral reinforcement effects have either restricted aggregate week-level data to a single domain (Villanueva et al. 2008) or published correlations without accounting for other variables such as customer satisfaction (Uncles et al. 2013). Others have focused only on incentivized referrals such that a referral reinforcement effect due to the reward could not be ruled out (Viswanathan et al. 2018). To date, we know little about individual-level effects tested across different settings and controlled for other potential drivers of referral behavior.

The primary aim of our research is thus to quantify referral reinforcement effects at the individual level, across domains and contexts. A second aim is to explore potential mechanisms that contribute to the referral reinforcement effect. It is reasonable to expect that customer satisfaction ranks high among customers who have been referred versus those not referred due to preference-matching and social enrichment between referral maker and recipient (Schmitt et al. 2011; Van den Bulte et al. 2018). In addition, higher satisfaction levels tend to favor the inclination to refer others in turn (Anderson 1998; De Matos and Rossi 2008). The next section provides more detail on this. Yet, empirical evidence of satisfaction's mediating role between receiving and making referrals is remarkably unavailable. Moreover, a critical question is whether customer satisfaction fully or partially

accounts for the referral reinforcement effect. If referral reception prompts the likelihood of extending a referral at least partly independent of satisfaction levels, then referrals would contribute to firm growth even more. Well accepted is the view "[i]n fact, the best source of new business is a referral from a satisfied customer" (Inc. 2010). Our research does not dispute this, and it explores the extent to which satisfaction is the key factor driving referral reinforcement. Yet, if referral reinforcement effects are sizeable but a substantial part of the effects is unmediated by satisfaction, encouraging referrals even when the satisfaction of the recipient is not maximal can still be profitable. We conducted four studies to examine these critical issues.

Study 1 is a field experiment among about 200,000 customers of a ridesharing platform. In support of a referral reinforcement effect, referred ridesharers tended to refer the service to others more versus those not referred. Further, the referral reinforcement effect was four times that of a firm's marketing intervention to stimulate referrals (10% vs. 7%) above a baseline rate (6%). Studies 2a and 2b explore the mediating role of customer satisfaction and decompose the referral reinforcement effect into satisfaction- versus non-satisfaction-mediated parts. Study 2a reanalyzes published transactional and survey data of a retail bank RRP. Study 2b enlists a new sample of U.S. moviegoers and controls for various drivers underlying referral receipt and retransmission. The satisfaction-mediated part was statistically significant but accounted for less than half (40% in Study 2a, and 43% in Study 2b) of the total referral reinforcement effect. Importantly, 60% and 57% accounted for non-satisfaction-mediated parts. Study 3 is a controlled lab-experiment about viewing television ads designed to rule out self-selection effects on the likelihood of making and receiving referrals. Finally, Study 4 is an experiment that explores consumer beliefs and motives to extend referrals. It finds that referral reception tends to amplify people's concern for others, motivating the referral gesture to others.

Next, we outline research that has probed the referral reinforcement effect and describe our conceptual framework. Then, we present our studies to quantify referral reinforcement effects across contexts and decompose them into satisfaction- versus non-satisfaction-mediated components. We conclude by discussing the implications of our findings.

2.2 The Referral Reinforcement Effect

Table 2.1 summarizes the literature that informs referral reinforcement effects. It focuses on studies that investigate the differences between referred and non-referred customers on any outcome. Early on, Sheth (1971) reported that referred U.S. customers of stainless steel razor blades showed higher referral rates than non-referred customers did. Likewise, German households who had recently switched energy providers due to a referral exhibited higher behavioral loyalty, which included making referrals, than those not referred (Von Wangenheim and Bayón 2004). In an effort to generalize, Uncles et al. (2013) found that referral rates across 15 product and service categories (such as supermarkets, dentists) were higher for customers disclosing *recommendation by others* rather than advertising as the main factor influencing their decision to purchase. These studies were based on observational data (column B in Table 2.1), and referrals emerged organically in the social network of existing and prospective customers without any firm promotion (column C). Using experimentally controlled referrals, Chen and Berger (2016) found people more likely to share high-quality online news articles after receiving these from others versus self-searched news items.

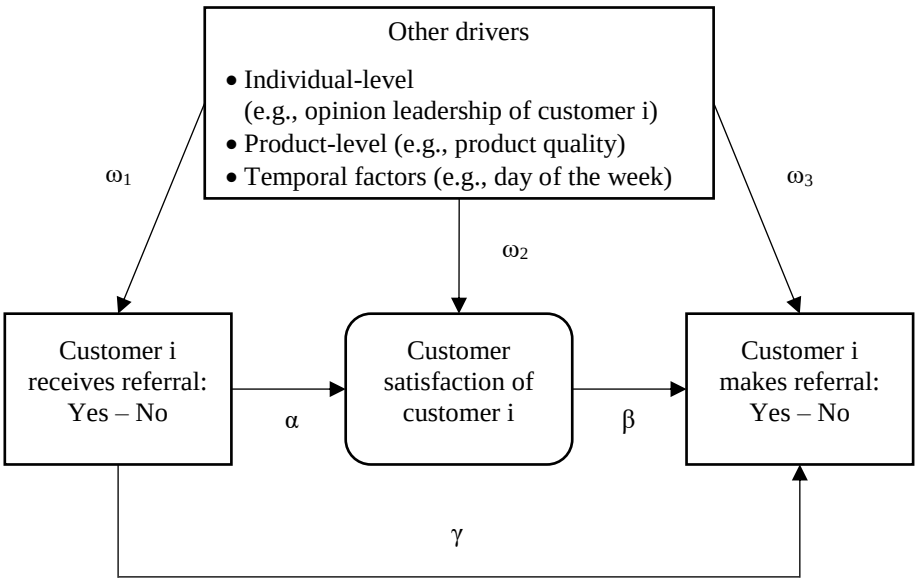
Figure 2.1 presents our conceptual framework. It specifies that customer satisfaction mediates the relationship between receiving and making a referral, and in addition that receiving a referral directly increases the likelihood of referral-making. It also includes other factors separately affecting the odds of receiving and making referrals. We discuss each pathway in turn.

Table 2.1
Referral Research Summary

(A) Study	(B) Setting	(C) Context of referrals	(D) Referral reinforcement effect	(E) Method	(F) Unit of analysis	(G) Control for satisfaction	(H) Findings: referred vs. non-referred customers
Sheth (1971)	601 U.S. razor blade users	Organic	Yes	χ^2-test: referred × referring	Customer	No	14%-points more likely to refer
Von Wangenheim and Bayón (2004)	367 German households that recently switched energy providers	Organic	Yes	χ^2-test: referred × referring	Household	No	15%-points more likely to refer
Verhoef and Donkers (2005)	3,317 customers of a Dutch financial-services provider	Organic	No (retention, cross-buying)	Probit model: retention & cross-buying = f(referred)	Customer	No	Lower retention rates (2/4 products) and more cross-buying (1/4 products)
Villanueva et al. (2008)	A free web hosting service	Organic	Yes	Vector autoregression (VAR): logins, customers acquired by marketing, referred customers	Week	No	3.6 vs. 1.8 new customer acquisitions
Trusov et al. (2009)	A social networking site	Organic	No (new sign-ups)	Vector autoregression (VAR): new sign-ups, referrals, traditional marketing	Week	No	Elasticity of sign-ups with respect to referrals is .53
Schmitt et al. (2011) & Van den Bulte et al. (2018)	9,495 customers of a German bank	Incentivized	No (contribution margin, retention, lifetime value)	Regression & hazard models: contribution margin, retention, lifetime value = f(referred)	Customer	No	25% higher contribution margin, 18% higher retention, 16% higher lifetime value
Uncles et al. (2013)	6,578 customers in 15 categories (such as supermarkets, dentists)	Organic	Yes	χ^2-test: referred × referring	Customer	No	11% more referrals made
Chen and Berger (2016)	6 experiments among 898 MTurkers & undergraduates	Manipulated	No (sharing)	ANOVAs: sharing = f(finding vs. receiving, high-quality vs. low-quality)	Participant	Yes (quality was manipulated)	Receivers share high-quality (vs. low-quality) content more but this difference is smaller for finders
Viswanathan et al. (2018)	796 customers of a financial services firm	Incentivized	Yes	Bivariate Poisson model: # referrals, # successful referrals = f(referred, satisfaction)	Customer	Yes (measured)	No difference in # referrals made (r = .05), more successful referrals made (r = .15)
Lee et al. (2018)	986 customers of a web annotation service, and 1,200 customers of an online file storage service	Incentivized & organic	No (usage of the service)	Hidden Markov model: usage = f(referred)	Customer	No	Less (more) usage of the web annotation (file storage) service
Present research	200,098 ridesharing customers, 470 customers of a bank, 810 moviegoers, 87 undergraduates, and 1,210 MTurkers	Incentivized, organic & manipulated	Yes	Path models and structural equation models: referring = f(referred, satisfaction)	Customer	Yes (measured and manipulated)	Referral reinforcement effect: referred customers refer more (r = .28, weighted r = .20)

Note: Table contains referral research, studies that investigated the differences between referred and non-referred customers on any outcome are included.

Figure 2.1
Framework for Referral Reinforcement Effects



Notes: Inclination to refer for customer i as a function of their satisfaction and a referral received (or not). Then, $\alpha*\beta$ is the satisfaction-mediated referral reinforcement effect, γ reflects the non-satisfaction-mediated referral reinforcement effect, and $\alpha*\beta+\gamma$ comprises the total referral reinforcement effect. The ωs represent the effects of other drivers that may account for the referral reinforcement effect.

2.2.1 Preference matching, social enrichment and customer satisfaction.

Customers who have been referred to a product or service are likely to be more satisfied with it than non-referred customers (Anderson 1998; De Matos and Rossi 2008). Figure 2.1 displays this. One reason is that referrers, unlike firms, are informed matchmakers. Referrers know their friends and acquaintances and are motivated to match them to the "right" product (Uncles et al. 2013). This improves preference matching by means of a more reasoned process of triadic balancing (the friend and product or service that one likes tend to be favorable to each other) and through more passive homophily where referrers recommend others similar to themselves (Schmitt et al. 2011; Van den Bulte et al. 2018). Recipients of referrals may even expect matchmaking to take place. The referral "tag" or "cue" might then be on itself sufficient to strengthen satisfaction (Hartline and Jones 1996) or referral receivers might attribute matching motives to the referrer (Verlegh et al. 2013). Social confirmation bias may also arise if the referral becomes the lens through which the referred brand is

experienced. In addition, customers receiving referrals are also likely to derive more value from a product than others owing to a mechanism of social enrichment (Schmitt et al. 2011). By following up referrals, positive experiences and feelings add to the shared history of the ones in the referral chain, deepening the bonds among them and the service or product.

In this way, these preference matching and social enrichment processes elevate post-consumption satisfaction in the referred customer which, in turn, could raise the inclination to make referrals (Anderson 1998; De Matos and Rossi 2008). For instance, customers make referrals to share their own satisfaction, to obtain positive recognition or praise, or to help others make the "right choice." Referring customers to a product or service that one enjoys and knows that others will like may also reinforce a consumer's bond with that product or service (Berger 2014).

2.2.2 Referral reinforcement independent of customer satisfaction.

It is reasonable to expect that receiving a referral may also raise the inclination to extend a referral *independent* of the satisfaction-mediated effect. First, referral reception may activate other-directed motives, such as a desire or moral duty to assist, indirect or generalized reciprocation (Baker and Bulkley 2014), or to generally do good to others (Campbell and Winterich 2018; Sundaram et al. 1998). Second, the person making the referral, and the endorsement itself, might signal social proof to the recipient for referring the product or service further (Chen and Berger 2016). A referral signals to the recipient that the product is being referred in the marketplace, or make referrals salient, which may in and of itself prompt further referrals regardless of the satisfaction level. Third, referral information can be personal and impassioned (Berger 2014), making the memory of consumption persist. When an opportunity later presents itself to make a referral to others, customers may then be more inclined to refer products or services that they readily remember.

The framework in Figure 2.1 decomposes the total effect of a referred customer passing forward the referral into two paths: satisfaction- versus non-satisfaction-mediated referrals. The magnitude of the satisfaction-mediated effect is cast as the product of the path from referral reception to satisfaction (α) times the path from satisfaction to referral extension (β). What remains is the direct non-satisfaction-mediated effect from receiving to making a referral (γ). If the referral reinforcement effect proved to be fully mediated by satisfaction, then firms would be well advised to focus on raising satisfaction levels in referral programs. In particular, they would have to be cautious about using referral rewards that incentivize customers to refer without paying attention to their potential satisfaction. Some (monetary) incentives (Ahrens et al. 2013; Jin and Huang 2014) are known to accentuate untargeted referral behavior that depresses the receiver's response to the referral (Verlegh et al. 2013). In contrast, the existence of a non-satisfaction mediated path would suggest that (high levels of) customer satisfaction need not be a condition for a referral reinforcement effect to occur. Thus, encouraging customers to refer *regardless of recipient satisfaction* could work simply since being referred *per se* activates repetition of the gesture.

The potential mediating role of customer satisfaction in converting referral receivers into referral makers (Table 2.1, column G) and the direct effect of referral reception untied to customer satisfaction have been largely unexplored. One exception controlled for customer satisfaction without proceeding to distinguish the satisfaction- versus non-satisfaction-mediated paths (Viswanathan et al. 2018).

Our framework accounts for other referral and satisfaction drivers. More specifically, individual differences such as age or gender (Kumar et al. 2010), and traits such as opinion leadership, can raise the likelihood of receiving and extending a referral (Iyengar et al. 2011). Product-level differences, such as product popularity, can also increase satisfaction to form suitable conversation topics, boosting the likelihood that a customer receives or makes a

referral (Berger 2014). Finally, temporal factors, even the day of the week, can influence referral reinforcement when referrals are received or made during certain times of the week.

2.2.3 Predictions and studies.

In sum, we predict that referred customers are more inclined to refer a product or service versus non-referred customers, and that this effect holds across industries and for both firm-incentivized and organic referrals. We expect customer satisfaction to mediate the referral reinforcement effect. We also expect that, circumventing the customer satisfaction route, referral reception increases the likelihood of its extension to others, while controlling for variables that could separately influence satisfaction, referral-making and receiving.

We present four studies to establish the referral reinforcement effect and explore its mechanisms. Our studies assess referral reinforcement for ridesharing (Study 1), retail banking (Study 2a), movie watching (Study 2b) and television commercials (Study 3). These studies enlist large-scale field data (Study 1), a combination of survey and archival data (Studies 2a and 2b), a controlled lab-experiment (Study 3), and a survey of customer beliefs about referral motives (Study 4). Studies 1 and 2a examine rewarded or incentivized referrals and use actual referral behaviors, while the remaining studies investigate organic referrals and measure referral intentions.

2.3 Study 1: Referral Reinforcement Effects among Ridesharing Customers

Study 1 establishes the presence and magnitude of a referral reinforcement effect in a field experiment among 200,098 customers of a ridesharing platform that featured its RRP. For successful referrals, this program rewarded both the referring and referred customers with $10 worth of credit toward their next rides. During the experiment, the ridesharing platform promoted its RRP to a random sample of customers, allowing us to compare the magnitudes of the referral reinforcement effect with the firm's intervention. Treated customers received a push notification (within 10 minutes after requesting the ride) touting the benefits of the

referral reward. The vast majority of rides charged the same cost. Since customer satisfaction was not directly measured, past usage variables served as proxies (Downing 1999). Ensuing studies contain direct measures of satisfaction and examined incentivized versus organic referrals. This dataset was used by Cohen et al. (2019) to examine the overall effectiveness of the push notifications but did not focus on the referral reinforcement effect.

2.3.1 Data and model.

Customers taking their second, third or fourth ride before the start of the intervention were included in the experiment. The analysis sample has 10,865 randomly selected "treated" customers (push notification) versus 189,233 non-treated customers. All these customers were similar users as to riding in the same city and completing their second, third or fourth ride before the start of the experiment. We observed whether a customer received a referral or not (REFERRED: 1 (Yes) or -1 (No)) and whether a customer was in the treatment or control group (TREATED: 1 (Yes) or -1 (No)). Our focal outcome for both the treatment and control groups is whether a customer makes a referral (REFERRING: 1 (Yes) or 0 (No)) within one week after the treatment. The dataset contains information on past usage to proxy customer satisfaction (Downing 1999): the number of past rides (PAST_RIDES), weeks since last ride (RECENCY), and weeks since user account creation (TENURE). These variables were standardized prior to the analyses. Appendix 2A presents summary statistics and code.

We estimated a binary Probit model to predict the probability that a customer refers (REFERRING) in the week following the intervention. The model for customer i was:

$$\begin{aligned} \mathrm{P}(\mathrm{REFERRING_i} = 1) = \Phi(&\omega_0 + \gamma \mathrm{REFERRED_i} + \omega_1 \mathrm{TREATED_i} + \\ &\omega_2 \mathrm{REFERRED_i} \times \mathrm{TREATED_i} + \beta_1 \mathrm{PAST\_RIDES_i} + \beta_2 \mathrm{RECENCY_i} + \\ &\beta_3 \mathrm{TENURE_i} + \omega_3 \mathrm{d_i} + \omega_4 \mathrm{h_i} + \zeta_i), \end{aligned} \tag{A2.1}$$

where Φ is the cumulative normal density, ω_0 is an intercept with regression parameters ωs, γ, and βs to be estimated, and where $\mathrm{d_i}$ and $\mathrm{h_i}$ represent day of the week (6 dummies, base is

Monday) and hour of the day (23 dummies, base is midnight) as fixed effects to rule out temporal determinants of the referral reinforcement effect. Lastly, $\zeta \sim N(0,1)$ is the error term.

We estimated two versions of the model. Model 1 omits effects of past usage (β_{1-3}) while Model 2 includes them. The focal γ quantifies the difference in the propensity to refer for referred versus non-referred customers to quantify *the referral reinforcement effect*. Further, we estimated the treatment effect of the push notification (ω_1) and investigated whether referred and non-referred customers respond differently to the treatment (ω_2). Significant negative interaction would imply that promoting the RRP curbs or even nullifies the referral reinforcement effect. This would imply that referred customers refer more merely due to awareness of the RRP and its monetary prize (they benefited from its reward already).

2.3.2 Results and discussion.

Table 2.2 presents the results. First, referral reception increases customer inclination to refer others within one week (tetrachoric correlation between the two binary variables = .18, $p < .001$). This referral reinforcement effect remains robust when controlling for day of the week and hour of the day fixed effects (Model 1: $\gamma = .14$, $p < .001$) and various proxies of satisfaction (Model 2: $\gamma = .14$, $p < .001$). Second, the marketing intervention yielded a positive effect on the inclination to refer ($\omega_1 = .04$, $p < .001$). Customers showed a baseline probability near 6% of referring, which then increased to 10% for referred customers versus

Table 2.2
Study 1: Referral Reinforcement Effects among Ridesharing Customers (n = 200,098)

		Model 1			Model 2		
Variable	Parameter	Estimate	SE	*P*-value	Estimate	SE	*P*-value
Intercept	ω_0	-1.761	(.034)	<.001	-1.762	(.034)	<.001
Receives referral (REFERRED)	γ	.136	(.011)	<.001	.142	(.011)	<.001
Receives treatment (TREATED)	ω_1	.024	(.011)	.024	.042	(.011)	<.001
REFERRED × TREATED	ω_2	-.014	(.011)	.189	-.014	(.011)	.184
# of past rides (PAST_RIDES)	β_1				-.003	(.005)	.563
# of weeks since last ride (RECENCY)	β_2				-.107	(.008)	<.001
# of weeks since account creation (TENURE)	β_3				-.065	(.007)	<.001

Notes: Results are from a binary Probit model with REFERRING (makes referral) as dependent variable. Table entries are unstandardized parameter estimates, standard errors (SE), and two-tailed *p*-values. Both models contain day of the week and hour of the day fixed effects, omitted from the table for brevity. R^2 estimates are .024 for Model 1 and .045 for Model 2. Details in Appendix 2A.

only 7% for customers who received the promotion. Thus, the referral reinforcement effect offered a 4:1 improvement compared to the marketing intervention. Third, the reinforcement effect did not differ between treated and control customers (ω_2 = -.01, p = .18). In other words, the reinforcement effect is robust to promoting the RRP. Thus, knowledge or saliency of the program and its rewards does not explain the referral reinforcement effect. Estimates of satisfaction proxies have face validity: users riding more recently (β_2 = -.11, p < .001) and enrolling more recently (β_3 = -.07, p < .001) were more prone to refer other customers.

In sum, this field experiment unveils a significant referral reinforcement effect while accounting for past usage variables as proxies for customer satisfaction. Importantly, the referral reinforcement effect offers fourfold the effect yielded by an intervention to promote referrals. A follow-up analysis tested the interactions between REFERRED and the satisfaction proxies and found evidence for an interaction between REFERRED and TENURE (β = -.03, p < .001; all other p > .32). Referred customers who joined the platform recently had a higher likelihood to refer, possibly due to the higher salience of the referral. Yet, a key limitation of this study is the unavailability of direct measures of customer satisfaction and individual-level characteristics that may account for joint variation in receiving and making referrals. Also, referrals were incentivized and limited to a specific ridesharing platform. The next studies address these very issues.

2.4 Study 2: Referral Reinforcement Effects and the Role of Satisfaction

Study 2 investigates two different settings, including referral incentivized and non-incentivized contexts, using direct measures of user satisfaction. Study 2a reanalyzes one published dataset (Ramaseshan et al. 2017) blending self-reported and archival data from a bank's RRP. Study 2b is a large-scale survey of moviegoers merged with archival data from a film database. It investigates referral reinforcement when referrals are organic, not incentivized.

2.4.1 Study 2a: Referral reinforcement effects among customers of a retail bank.

Data and model.

Ramaseshan et al. (2017) merged survey data of 470 customers of an international retail bank with transaction data of its RRP and reported summary statistics. The RRP offered a reward, such as a coffeemaker, to customers who referred others to become paying customers of the bank. Transaction data indicated that half of the 470 customers were referred while the other half were acquired by other means (REFERRED: 1 (Yes) or 0 (No)). The satisfaction measure (SAT) featured two items (Cronbach's α reliability = .85): "Bank X absolutely fulfills my expectations" and "Overall, I'm very satisfied with Bank X." Both responses scored from 1 (strongly disagree) to 7 (strongly agree). Five items (α reliability = .94) captured referral behavior (REFERRING) using the same response scale: "I often recommend Bank X", "I often recommend Bank X to close relatives and friends", "I often recommend Bank X to colleagues and acquaintances", "I often recommend Bank X when somebody is asking me about related advice" and "I often tell positive things about Bank X when I am asked."

Table 2.3
Study 2a: Referral Reinforcement Effects among Customers of a Retail Bank (n = 470)

Variable	Parameter	Estimate	SE	*P*-value
Customer satisfaction (SAT)				
Receives referral (REFERRED)	α	.414	(.104)	<.001
Makes referral (REFERRING)				
Customer satisfaction (SAT)	β	1.008	(.074)	<.001
Receives referral (REFERRED)	γ	.624	(.143)	<.001
Referral reinforcement effect decomposition	Parameter	Estimate	95% CI	
SAT-mediated effect	$\alpha*\beta$	.406	[.120, .706]	
Non-SAT-mediated effect	γ	.634	[.255, 1.014]	
Total referral reinforcement effect	$\alpha*\beta+\gamma$	1.040	[.594, 1.483]	
% SAT-mediated effect	$(\alpha*\beta)/(\alpha*\beta+\gamma)$	40%		
% non-SAT-mediated effect	$\gamma/(\alpha*\beta+\gamma)$	60%		

Notes: Table entries for top panel are unstandardized parameter estimates, standard errors (SE), and two-tailed *p*-values from a single-indicator structural equation model. The bottom panel lists mean estimates and 95% CIs based on 25,000 Monte Carlo replications. R^2 of SAT was .038, and R^2 of REFERRING was .421. Details in Appendix 2B.

We estimated a single-indicator structural equation model (SI-SEM) to quantify the referral reinforcement effect and its extent mediated by satisfaction. The SI-SEM generalizes standard regression and path models, which assume that predictors are measured without error and which lead to biased estimates if the assumption is violated, to situations where information about measurement error of predictors is available. Here such information is available because measurement error is 1-reliability, and Cronbach's alpha is an estimate of reliability. Mediation analyses rarely correct for measurement error, which can lead to severely biased estimates of indirect and direct effects (Pieters 2017). Appendix 2B provides further details and the code.

Results.

Table 2.3 reports the results. First, there is a sizeable effect of referral-receiving toward referral-making (point-biserial correlation corrected for attenuation = .29, $p < .001$). Second, the satisfaction-mediated referral reinforcement effect proves statistically significant ($\alpha*\beta$ = .41, 95% CI [.12, .71]), meaning that being referred increases customer satisfaction (α = .41, $p < .001$), and that satisfied customers are more inclined to refer (β = 1.01, $p < .001$). While substantial, the satisfaction-mediated effect accounts for only 40% of the total effect. The remaining 60% bypasses the satisfaction route (γ = .62, $p < .001$).

2.4.2 Study 2b: Referral reinforcement effects among moviegoers.

Study 2b is a large survey of moviegoers merged with IMDb data on movie quality ratings and gross earnings. It investigates referral reinforcement effects in a setting where customers made referrals organically without a firm's intervention. Study 2b controls for various movie- and individual-level drivers of the referral reinforcement effect to minimize the likelihood of omitted variable bias detailed below.

Data and measurement.

Nine hundred U.S. MTurk participants completed a survey on movie consumption. Participants were included when they had seen a movie in a theater during the past 12 months. Participants disclosed the movie title and answered a set of questions about the experience. We merged the survey data with movie-level data from IMDb. Responses for movies with missing IMDb data or having duplicate IP addresses were excluded. The final sample comprised 851 participants (509 females, mean age = 32).

Participants disclosed two items: whether they had received a referral to see the movie (REFERRED: 1 (Yes) or 0 (No)): "Did anyone recommend this movie before you saw it?" and whether they had already made or planned referrals to others for this specific movie (REFERRING: 1 (Yes) or 0 (No)) (Brown et al. 2005). About 95% of the referrals came from a partner, family member, and/or friend. Three items (Maxham III and Netemeyer 2002) assessed participant movie satisfaction (SAT), including "I am satisfied with my overall experience with the movie" using a seven-point scale 1 (strongly disagree) to 7 (strongly agree). Its composite reliability (CR) was .77 per confirmatory factor analysis. Five items with the same response scales assessed opinion seeking (SEEK), including "I like to get others' opinions before I see a movie" (CR = .88). Six items adopted from Flynn et al. (1996) enlisting the same seven-point scale assessed opinion leadership directly (LEADER), including "I often persuade other people to see movies that I like" (CR = .87). In addition, participants indicated gender (GENDER: 1 (Male) and 0 (Female)) and age (AGE, in years). For all mentioned movies, we added the following information from IMDb: opening weekend box office revenue (BOX: natural logarithm of U.S. dollars) and Metascore quality rating (RATE: 0-100). Appendix 2C presents measurement-estimation details and the code.

Model.

We specified a generalized structural equation model (GSEM) to handle the binary variables REFERRED and REFERRING while correcting for measurement error in latent variables and controlling for other potential drivers of the referral reinforcement effect (Figure 2.1). For instance, opinion leaders (LEAD) tend to refer others independent of the specific movie (Iyengar et al. 2011). Likewise, individuals with a higher propensity to seek advice (SEEK) tend to receive referrals independent of the movie title (Flynn et al. 1996). We controlled for gender and age since customers with certain demographic profiles may refer or rely more on referrals (Kumar et al. 2010). Because popular and blockbuster movies generally raise the probability of referrals, we controlled for a movie's opening weekend box office revenue (BOX) and rating (RATE). The structural model is:

$$\mathrm{P}\left(\mathrm{REFERRED}_{1,\mathrm{i}}=1\right)=\Phi\left(\sum_{\mathrm{k}=0}^{6}\omega_{1,\mathrm{k}}\mathrm{COV}_{\mathrm{k,i}}+\zeta_{1,\mathrm{i}}\right), \tag{2.2}$$

$$\mathrm{SAT}_{2,\mathrm{i}}=\alpha\mathrm{REFERRED}_{\mathrm{i}}+\sum_{\mathrm{k}=0}^{6}\omega_{2,\mathrm{k}}\mathrm{COV}_{\mathrm{k,i}}+\zeta_{2,\mathrm{i}}, \tag{2.3}$$

$$\mathrm{P}\left(\mathrm{REFERRING}_{3,\mathrm{i}}=1\right)= \Phi\left(\beta\mathrm{SAT}_{\mathrm{i}}+\gamma\mathrm{REFERRED}_{\mathrm{i}}+\sum_{\mathrm{k}=0}^{6}\omega_{3,\mathrm{k}}\mathrm{COV}_{\mathrm{k,i}}+\zeta_{3,\mathrm{i}}\right), \tag{2.4}$$

where Φ is the cumulative normal density (Equations 2.2 and 2.4 are binary Probit regressions), ωs, α, β and γ are regression parameters, and $\zeta\sim\mathrm{N}(0,\sigma_{\zeta}^{2})$ comprises the iid error terms. Covariates COV consist of an intercept, SEEK, LEADER, GENDER, AGE, BOX, and RATE. Since the model complexity prevented standard ML estimation, we used Bayesian estimation (with 25,000 MCMC iterations and default non-informative priors) to estimate parameters. Details and annotated code appear in Appendix 2C.

Results.

Models 1 and 2 without and with covariates (COV), respectively, yielded very similar results, which is reassuring. Table 2.4 reports estimation results for Model 1 (details for both models are in Appendix 2C). The results converge with those of Studies 1 and 2a. First, there is clear evidence of referral reinforcement effect: the association between referral-receiving and -making is statistically significant and substantial (tetrachoric correlation between the two binary variables = .38, $p < .001$). Second, the referral reinforcement effect is mediated by satisfaction (Model 1: $\Phi(\alpha*\beta) = .08$, 95% CI [.05, .12]). Movie satisfaction rises for customers who had been referred versus those non-referred ($\alpha = .42$, $p < .001$) with satisfied customers more likely to refer others to the movie they watched ($\beta = .92$, $p < .001$). Third, although the satisfaction-mediated effect is sizeable, it accounts for only 43% of the total referral reinforcement effect. The non-satisfaction-mediated effect accounts for 57% of the total effect ($\Phi(\gamma) = .11$, 95% CI [.05, .17]). The size of these effects is not biased by measurement error in satisfaction since that was accounted for by the model.

Table 2.4
Study 2b: Referral Reinforcement Effects among Moviegoers (n = 851)

Variable	Parameter	Estimate	SD	*P*-value
Movie satisfaction (SAT)				
Intercept	$\omega_{2,0}$	5.324	(.151)	<.001
Receives referral (REFERRED)	α	.423	(.075)	<.001
Makes referral (REFERRING)				
Intercept	$\omega_{3,0}$	-4.465	(.411)	<.001
Movie satisfaction (SAT)	β	.922	(.079)	<.001
Receives referral (REFERRED)	γ	.440	(.120)	<.001
Referral reinforcement effect decomposition	Parameter	Estimate	95% CI	
SAT-mediated reinforcement effect	$\Phi(\alpha*\beta)$	.083	[.053, .116]	
Non-SAT-mediated reinforcement effect	$\Phi(\gamma)$	.114	[.054, .174]	
Total reinforcement effect	$\Phi(\alpha*\beta+\gamma)$	.197	[.137, .255]	
% SAT-mediated reinforcement effect	$\Phi((\alpha*\beta)/(\alpha*\beta+\gamma))$	43%		
% non-SAT-mediated reinforcement effect	$\Phi(\gamma/(\alpha*\beta+\gamma))$	57%		

Notes: Latent variables identified by fixing variance to unity. Estimates of regression and path weights are unstandardized, posterior standard deviations (SD) and one-tailed Bayesian *p*-values. The referral reinforcement effect decomposition gives estimates and 95% CIs. $\Phi(\cdot)$ denotes effects on the latent response variable back-transformed along the Probit probability curve. R^2 estimates are .042 for SAT and .505 for REFERRING. Details in Appendix 2C.

The referral reinforcement effect is robust when controlled for other determinants of getting or making referrals and for customer satisfaction (Model 2). The sign and size of the effects of these covariates are as expected. For instance, opinion seekers are more likely to be referred ($\omega_{1,1}$ = .30, p < .001) while opinion leaders are more satisfied ($\omega_{2,2}$ = .20, p < .001) and more likely to refer ($\omega_{3,2}$ = .47, p < .001). Also, higher rated movies ($\omega_{1,4}$ = .04, p < .001) are more prone to referral reception and elevate levels of satisfaction ($\omega_{2,4}$ = .03, p < .001). Together, the covariates increase the variance accounted for by the predictors in satisfaction (R^2 from .04 in Model 1 to .19 in Model 2) and in referring (from .51 to .61). The size of the referral reinforcement effect drops from .197 (Model 1) to .085 (Model 2).

Importantly and similar to Model 1, 43% of the referral reinforcement effect remains satisfaction-mediated ($\Phi(\alpha*\beta)$ = .035, 95% CI [.01, .06]) while 57% is non-satisfaction-mediated ($\Phi(\gamma)$ = .05, 95% CI [.01, .09] in Model 2. Thus, compared to non-referred customers, referred customers have on average about an 8.5 percentage-point higher probability of referring others versus those non-referred, where about 3.5 percentage-points (43%) ascribe to satisfaction-mediated referral reinforcement, leaving 5 points (57%) for non-satisfaction-mediated referral reinforcement.

Follow-up analyses ruled out an interaction effect between satisfaction and referral reception on referral making. Importantly, the interaction between satisfaction and referral reception did not significantly affect likelihood to refer others (β = -.15, p = .17) beyond the two main effects. Details and supplemental robustness checks are in Appendix 2C.

2.4.3 Discussion.

Study 2 decomposed the referral reinforcement effect into satisfaction- versus non-satisfaction-mediated parts. The non-satisfaction-mediated part accounts for about 60% of the total referral reinforcement effect in both sub-studies. Study 2 also generalized the referral reinforcement effect from Study 1 (ridesharing) across different industries (Study 2a: retail

banking, and Study 2b: movies) and beyond the context of a RRP (Study 2b). Study 2b controlled for various relevant covariates. Yet, self-reported opinion leadership measures might capture self-confidence rather than actual influence captured by network-based measures, unavailable to us (Iyengar et al. 2011). In sum, likelihoods of omitted variables accounting for some of the referral reinforcement effect cannot be dismissed. Study 3 was thus conducted under controlled conditions.

2.5 Study 3: Referral Reinforcement Effects when Referring Commercials

Study 3 is a controlled lab-experiment crafted to rule out alternative explanations for the referral reinforcement effect and to extend the previous findings. First, Study 3 uses an experimental design that randomly assigns subjects to being referred (treatment) or not-referred (control). This rules out the possibility that the same factors influence the likelihood of referral-making from referral-receiving. After random assignment, all participants experienced the same viewing event to then score their satisfaction levels with the event and inclinations to refer.

Second, random assignment rules out satisfaction-mediated referral reinforcement by preventing preference-matching to take place. Specifically, since all participants experienced the same event and assignment to the referral reception condition was random, referrers could not use matchmaking ability to recommend the "right" product to the "right" customer. In real life, people tend to refer when they expect the recipient to likely enjoy the product (Van den Bulte et al. 2018). Preventing better matching from taking place allows us to focus on the remaining referral reinforcement effect while still controlling for differences in participant satisfaction after the viewing event. Thus, we expect that the referred manipulation does not affect satisfaction while yielding a sizeable referral reinforcement effect, even after controlling for differences in satisfaction.

Third, Study 3 uses broader measures of satisfaction to rule out the possibility that modest satisfaction-mediated effects in Study 2 are due to the specific satisfaction measures. The satisfaction measures in Studies 2a and 2b are common, but are more cognitive than affective in nature, and this could lead to underestimating the satisfaction-mediated reinforcement effects. Affective evaluation is more spontaneous and automatic than cognitive evaluation which is more conscious and deliberate, and both types can elicit distinct effects (Wilcox et al. 2011). Results of a follow-up analysis in Study 2b (Appendix 2C) using a squared satisfaction term to capture effects of extreme satisfaction levels makes it unlikely that failure to capture affective response accounts for the modest satisfaction-mediated effect (Anderson 1998). Still, isolating the cognitive and affective measures of satisfaction in Study 3 further rules out such a possibility.

In this setting, we predict three effects: (1) being referred increases the inclination toward referral-making, (2) being referred versus non-referred does not influence satisfaction, and (3) the reinforcement effect remains intact after controlling for differences in customer satisfaction measured more comprehensively.

2.5.1 Participants, design and procedure.

Eighty-seven paid undergraduate students (56 females, 31 males, mean age = 20) engaged the behavioral lab in dyads. We invited pairs to ensure that "referred" manipulation was relevant and convincing to the participants. A four-item tie-strength measure (CR = .73) from Ryu and Feick (2007) assessed the strength of each tie. Dyads had strong ties as reflected by the high mean of the first item (M = 7.26, SD = 1.94) on an ten-point scale and that of the remaining three items (M = 4.94, SD = .90) on six-point scales (Ryu and Feick 2007). Members of dyads (and one triad) were separated, and each participant was randomly assigned to a condition (referred or control) by an experimenter blind to the predictions. No participants were dropped.

After several unrelated studies, all participants were told that they were about to watch a television commercial. Participants in the referred condition were instructed to exit their cubicles and collect a set of headphones from the experimenter. The experimenter handed out the headphones and administered the manipulation. Participants were asked for the name of the person with whom they came to the lab. After a brief pause, the experimenter indicated that this person had already finished the study, watched the commercial, and had left a brief personal message stating that the other person liked the commercial and recommended it. Then, participants were escorted back to their cubicles and instructed to wear headphones and view a 2015 Super Bowl commercial called "Settle it" featuring fruit-flavored Skittles sweets. A familiarity check verified the commercial was unknown to all participants. Participants in the control (not-referred) condition watched the commercial wearing headphones provided when entering the cubicle. In reality, participants did not leave messages to each other after viewing the same commercial, and all subjects in the referred group heard the same message. This manipulation prevented preference-matching to identify the non-satisfaction-mediated referral reinforcement effect while controlling for differences in satisfaction.

2.5.2 Measurement and model.

After watching the commercial, seven items assessed affective evaluation (AFF) on an 11-point semantic differential scale using anchors "not enjoyable" versus "enjoyable", "boring" versus "interesting", "unpleasant" versus "pleasant", "unlikable" versus "likable", "depressing" versus "uplifting", "not entertaining" versus "entertaining", and "irritating" versus "not irritating" (Wilcox et al. 2011) (CR = .94). Cognitive evaluation (COG) was measured enlisting four items: overall quality of the commercial, quality of the acting, quality of the story, and quality of the production, all on 11-point scaling with anchors "poor" versus "excellent" (CR = .80). The order of COG and AFF was counterbalanced. Four items

assessed customer intentions to refer (REFERRING). Subjects scored whether they would like to: share this commercial with others, recommend others view the ad, speak positively, or speak negatively (reversed item) of "Settle it" in conversation using the 11-point scale with anchors "completely disagree" versus "completely agree" (Brown et al. 2005) (CR = .84).

We estimated a structural equation model regressing the two latent satisfaction variables on the "being referred" manipulation while regressing the latent "referring others" variable on the two satisfaction measures and the "being referred" manipulation. Bootstrapping 25,000 replications and the 95% CI assessed direct and indirect effects. Additional measurement details and the code appear in Appendix 2D.

2.5.3 Results and discussion.

Table 2.5 reports the results. As predicted, being referred exerts statistically significant impact on referring (point-biserial correlation corrected for attenuation = .26, $p = .02$), even while controlling for cognitive and affective satisfaction ($\gamma = .66$, $p = .04$), demonstrating once more a referral reinforcement effect. Further, as predicted, the "being referred"

Table 2.5
Study 3: Referral Reinforcement Effects when Referring Commercials (n = 87)

Variable	Parameter	Estimate	SE	*P*-value
Affective evaluation (AFF)				
Receives referral (REFERRED)	α_1	.196	(.220)	.373
Cognitive evaluation (COG)				
Receives referral (REFERRED)	α_2	.150	(.240)	.532
Makes referral (REFERRING)				
Affective evaluation (AFF)	β_1	.891	(.284)	.002
Cognitive evaluation (COG)	β_2	1.046	(.357)	.003
Receives referral (REFERRED)	γ	.662	(.322)	.040

Referral reinforcement effect decomposition	Parameter	Estimate	95% CI
SAT-mediated reinforcement effect via AFF	$\alpha_1*\beta_1$	.175	[-.159, .932]
SAT-mediated reinforcement effect via COG	$\alpha_2*\beta_2$	.157	[-.298, 1.338]
Non-SAT-mediated reinforcement effect	γ	.662	[.021, 1.600]
Total reinforcement effect	$\alpha_1*\beta_1+\alpha_2*\beta_2+\gamma$	.994	[.014, 2.483]
% SAT-mediated reinforcement effect	$(\alpha_1*\beta_1+\alpha_2*\beta_2)/(\alpha_1*\beta_1+\alpha_2*\beta_2+\gamma)$	33%	
% non-SAT-mediated reinforcement effect	$\gamma/(\alpha_1*\beta_1+\alpha_2*\beta_2+\gamma)$	67%	

Notes: Latent variables identified by fixing variance to unity. Top panel lists unstandardized parameters estimates of regression and path weights with standard errors (SE) and two-tailed *p*-values. Estimates and 95% CI in the bottom panel are based on 25,000 bootstrapped samples. Estimated correlation between residuals of AFF and COG is .715. R^2 estimates for AFF, COG and REFERRING are .010, .006, and .776, respectively. Details in Appendix 2D.

manipulation did not influence the two measures of satisfaction from viewing the commercial (affective evaluation $\alpha_1 = .20$, $p = .37$, cognitive evaluation $\alpha_2 = .15$, $p = .53$), while both measures did influence inclinations to refer others (affective evaluation $\beta_1 = .89$, $p < .01$ and cognitive evaluation $\beta_2 = 1.05$, $p < .01$). Hence, there was no substantive satisfaction-mediated referral reinforcement ($\alpha_1*\beta_1 = .18$, 95% CI [-.16, .93]; $\alpha_2*\beta_2 = .16$, 95% CI [-.30, 1.34]). Further validating the referral reinforcement effect, the non-satisfaction mediated impact of being referred toward referring others accounts for 67% of the total effect.

In sum, Study 3 supported the referral reinforcement effect under controlled conditions. The results are consistent with the expectation that the manipulation averted subjects serving as matchmakers, to successfully block satisfaction-mediated referral reinforcement. Although other mechanisms such as a stronger attachment to the advertised product due to social enrichment (Schmitt et al. 2011) are not necessarily blocked by the manipulation, the results are consistent with a direct experience account. Evaluation of the commercial is unaffected by prior information, in this case a referral, because participants were able to directly and unambiguously evaluate the commercial (Hoch and Ha 1986).

2.6 Study 4: Exploring Customers' Lay Beliefs about Referral Motives.

A wide range of distinct motives and drivers, indicated in the theory section, can separately or jointly contribute to a referral reinforcement effect beyond customer satisfaction. Rather than test the impact of one or more specific referral motives, the objective of Study 4 was to explore, using a broad brush, lay beliefs that referred and non-referred customers held about these motives. Though customer beliefs may be at variance with the actual mechanisms (Friestad and Wright 1995), they nonetheless influence people's attitudes and decisions (McFerran and Mukhopadhyay 2013) and are instructive in and of themselves. The results of Study 4 can inform future theory and research about the referral reinforcement effect as later detailed in the Discussion.

2.6.1 Participants, design, procedure and measurement.

A sample of 1,251 U.S. MTurk participants completed a survey about making positive recommendations to others. We dropped 41 having duplicate IP addresses (final n = 1,210, 51% female, mean age = 40). Participants read a scenario describing a consumption situation and were asked to envision themselves in it. They were assigned randomly to a condition in a 2 (referred: referred vs. not referred) × 5 (satisfaction level: extremely dissatisfied, moderately dissatisfied, not dissatisfied or satisfied, moderately satisfied, extremely satisfied) between-participants experimental design. We varied satisfaction between participants to explore the possibility that lay beliefs about referral motives vary across satisfaction levels. The levels, ranging from extremely dissatisfied to extremely satisfied, allowed us to focus on a broad range of situations, also those that might be less common such as referrals by dissatisfied customers. Preference matching by dissatisfied customers might still occur when they are motivated to diffuse the product to the "right" people in their network. Yet, these situations might not be observed when participants are asked to recall experiences instead of manipulating situations (Bougie et al. 2003), due to demand effects or memory.

All participants were informed that a new product had been introduced and that they considered buying it. The specific product was undisclosed to avoid leading the participants. Those in the referred condition also read that a friend had recommended the product to them and that they had decided to buy it. Those in the not-referred condition read that no one had recommended them the product at the time of purchase. All participants imagined buying the product and consuming it. The scenario then manipulated user satisfaction level. To illustrate, participants in the extremely satisfied condition read that they really liked the product, being exceptionally happy and extremely satisfied with the product. All participants then read that they bought and consumed the product, met a friend, and decided to recommend the product. Manipulation checks indicated that the majority of participants correctly identified both their

referral status (93% of sample) and manipulated satisfaction level (90%). All participants were to imagine referring the product to another person.

Referral motives can be broadly classified into self-directed motives (showing one's expertise, seeking advice), other-directed motives (desire or moral obligation to help specific others or generally do good), and product-directed motives (boosting product success or venting one's excitement about it) (Alexandrov et al. 2013; Berger 2014; Bronner and De Hoog 2011; Engel et al. 1969; Sundaram et al. 1998). Subjects responded to each of 23 items to identify these categories and how likely each applied to them in their situation (Extremely unlikely (1) to Extremely likely (5)). Our approach aligns with that of Bougie et al. (2003), but we manipulated experiences beyond just sampling them. Eleven items covered self-directed motives such as self-enhancement ("It makes you look good") and sense-making ("You will learn from your friend's experience with this product"). Eight items covered other-directed motives such as a desire to help others ("It will help your friend decide whether to buy this product or not") or meeting a moral obligation to refer ("You feel obligated to do so"). Finally, four items covered product-directed motives ("You will help this product succeed"). Participants also disclosed gender and age.

A three-factor confirmatory factor analysis (3-CFA) of the 23 items revealed acceptable global ($\chi^2_{(227)}$ = 3,718; CFI = .79; RMSEA = .11; SRMR = .08) and good local fits (composite reliabilities of .90, .85, and .81 for self-, other- and product-directed factors, respectively). A one-factor CFA fitted the data worse, supporting discriminant validity of the three-factor solution. Appendix 2E presents sample scenarios, measurement details, and the code.

2.6.2 Model, results and discussion.

We performed a latent MANOVA using structural equation modeling with the three referral motive factors as dependent (latent) variables along with the referred manipulation

(REFERRED: -1 (not referred) or 1 (referred)), the satisfaction manipulation (SAT: -2 (extremely dissatisfied) to 2 (extremely satisfied)), and the interaction between SAT and REFERRED as manifest predictors. The error terms were free to covary.

Means and standard deviations are reported in Table 2.6, and detailed model estimates appear in Appendix 2E. First, a main effect from satisfaction level emerged: elevated levels of satisfaction increased all three motives (self-, other- and product-directed) to refer others to the new product (all *p*-values < .01). Second, participants who had been referred themselves were more likely to refer for *other-directed* motives versus those who had not been referred ($\gamma = .13$, $p < .01$). Referral-reception status, however, did not influence self-oriented ($\gamma = .03$, $p = .26$) or product-directed ($\gamma = -.01$, $p = .82$) motives to refer. This documents the lay belief that receiving a referral sensitizes people's concern for others and raises their intentions to help by passing on a referral. Third, an interaction effect between satisfaction and receiving a referral on other-directed motives emerged ($\beta = -.06$, $p < .01$), but not for self- and product-directed motives (*p*-values > .11). Post-hoc analyses indicate that heightened other-directed motives were strengthened in those dissatisfied ($p = .03$) or extremely dissatisfied ($p < .01$).

Table 2.6
Study 4: Customers' Lay Beliefs about Referral Motives (n = 1,210)

Condition: REFERRED	Condition: SAT					
	Extremely Dissatisfied	Dissatisfied	Neither/nor	Satisfied	Extremely Satisfied	Mean (SD)
			Self-directed motives			
No	1.99 (.76)	2.28 (.76)	2.67 (.71)	2.79 (.66)	2.48 (.66)	2.51 (.78)
Yes	2.17 (.80)	2.38 (.82)	2.58 (.82)	2.87 (.60)	2.81 (.74)	2.56 (.81)
Mean (SD)	2.08^{a} (.78)	2.33^{b} (.79)	2.63^{c} (.77)	2.83^{d} (.63)	2.83^{d} (.70)	
			Other-directed motives			
No	2.47 (.86)	2.69 (.82)	3.17 (.63)	3.47 (.50)	3.59 (.52)	3.07 (.81)
Yes	2.81 (.83)	2.89 (.78)	3.20 (.75)	3.54 (.47)	3.65 (.55)	3.22 (.77)
Mean (SD)	2.63^{a} (.86)	2.79^{b} (.80)	3.18^{c} (.69)	3.50^{d} (.49)	3.62^{d} (.53)	
			Product-directed motives			
No	2.16 (.88)	2.41 (.82)	2.93 (.78)	3.24 (.70)	3.48 (.70)	2.83 (.93)
Yes	2.36 (.88)	2.57 (.90)	2.82 (.92)	3.27 (.65)	3.43 (.75)	2.89 (.92)
Mean (SD)	2.26^{a} (.88)	2.48^{b} (.86)	2.88^{c} (.85)	3.26^{d} (.67)	3.46^{e} (.72)	

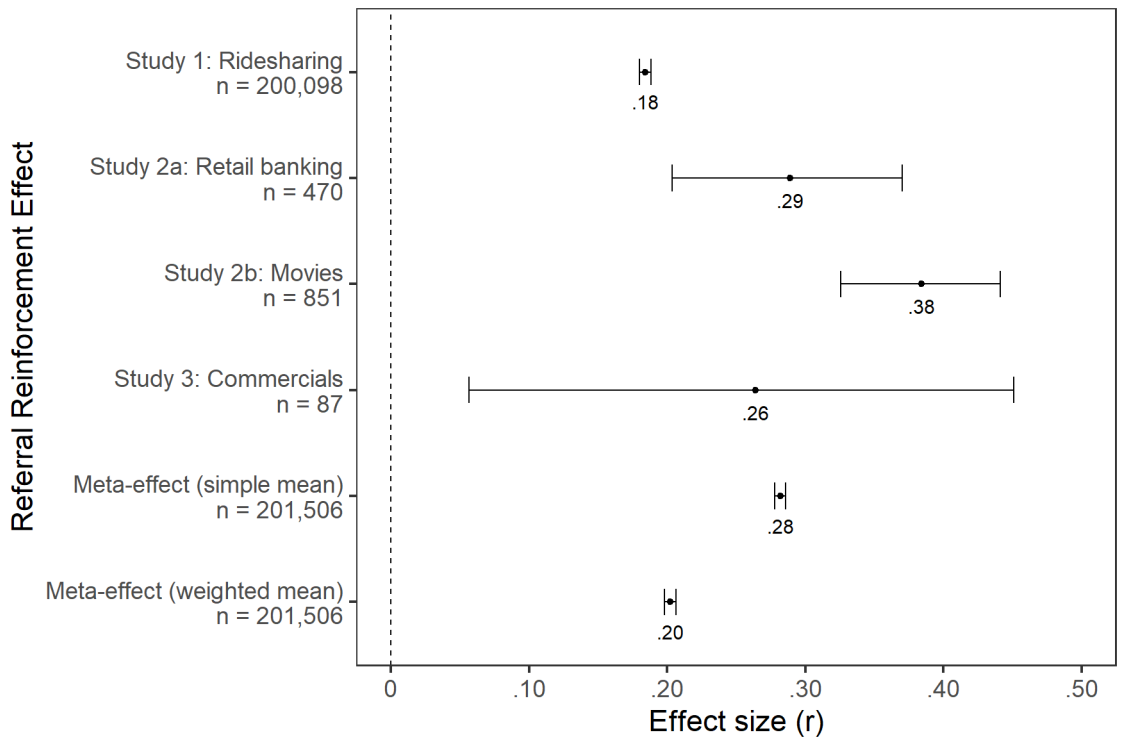
Notes: Latent means and standard deviations (SD) from a structural equation model are shown casting referral motives as three multi-item latent variables. Different superscripts row-wise indicate significant differences between respective column means at $p < .05$. Details in Appendix 2E.

Satisfied customers refer due to other-directed motives, regardless of having been referred or not, but dissatisfied customers are more motivated to refer if they were referred themselves. The referral might signal social proof (Chen and Berger 2016), and referred customers might be more motivated to match the product with a person who is a better fit than they themselves were (Schmitt et al. 2011; Van den Bulte et al. 2018). An alternative explanation is that the scenario where referred dissatisfied customers recommend is uncommon to participants, and recommending to benefit others was the only reason they could think of.

2.7 Discussion

First, we find consistent support for the presence and generality of a referral reinforcement effect: customers receiving referrals are more prone to refer that same product or service to others compared to those who did not. This referral reinforcement effect was found across industries, samples, methodologies, and for both incentivized and organic referrals. The

Figure 2.2
Meta Referral Reinforcement Effect



Notes: Forest plot has total referral reinforcement effect sizes, based on disattenuated correlations without accounting for covariates, with 95% CIs based on Fisher-Z transformations. Meta-effects are based on simple (unweighted) means and weighted means by standard error of Fisher-Z. Details in Appendix 2F.

different measures of referrals (both self-reported and actual referrals) rule out that common method bias fully accounts for the referral reinforcement effect. The forest plot in Figure 2.2 visualizes the total effect sizes from Studies 1-3, corrected for measurement error and without satisfaction or covariates accounted for. The size of the average referral reinforcement effect was r = .28 (Cohen's d = .58) based on simple means and r = .20 (Cohen's d = .41) based on weighted means across Studies 1-3. Details are in Appendix 2F. This referral reinforcement effect is sizeable given r = .20 is much larger than the r = .02 effect of a marketing intervention to stimulate referrals (Study 1). Accounting for covariates yielded a referral reinforcement meta-effect size of .19 (Cohen's d = .39).

Second, greater satisfaction among customers who had been referred accounted for nearly 40% of the referral reinforcement effect versus those not referred in our studies.

Third, while statistically or experimentally controlling for their satisfaction levels, customers who had been referred were more likely to refer in turn, accounting for 60% of the total effect in our studies. The non-satisfaction-mediated referral reinforcement effect was quantified for various operationalizations of the satisfaction construct, while controlling for measurement error and potential confounders and randomly assigning participants to the being-referred treatment (Study 3). Together, this supports the robustness of the referral reinforcement effect.

Fourth, our final Study 4 suggests that receiving a referral increases concern for others that motivates referred customers to extend the referral. This effect could partially account for the referral reinforcement effect among otherwise equally satisfied customers.

2.7.1 Implications for marketing theory and practice

The results have implications for marketing theory. First, they contribute to customer management theory. The benefits of referral reward programs (RRP) in recruiting high-CLV customers arising from preference matching and social enrichment are known (Schmitt et al.

2011; Trusov et al. 2009; Van den Bulte et al. 2018; Villanueva et al. 2008). However, the referral reinforcement effect implies that, in addition to CLV benefits, customers who have been referred also yield a larger customer referral value and are thus even more valuable to firms than generally considered. Notably, this benefit emerges beyond mere increased satisfaction in referred customers due to preference matching and social enrichment: most of the referral reinforcement effect (60%) was *independent* of recipient satisfaction. Our research also rules out that referral reinforcement effects are driven merely by greater salience or awareness of an impending referral reward: promoting the program to customers did not attenuate the referral reinforcement effect.

Second, our finding that lay beliefs about referral reinforcement favor other-directed motives offers a promising path for further inquiry. Research has recognized self- versus other-directed motives for referrals (Alexandrov et al. 2013; Bronner and De Hoog 2011). We speculate that increased concern for others among referred customers may prod them to extend referrals with altruistic motives or moral obligations to reciprocate in mind (Baker and Bulkley 2014), despite the self-serving nature of much word-of-mouth encounters (Berger 2014). Future work can test this theory.

The results also add to the literature on longer chains or cascades of social influence (Goel et al. 2015; Leskovec et al. 2007). Although we cannot infer higher-order cascades, our findings unveil potential impetus from referral reinforcement effects in fueling such cascades, especially indicating that referral reinforcement effects largely bypass satisfaction levels.

Furthermore, the findings have managerial implications for RRPs. First, the fact that referral reinforcement effects surpassed the impact of promoting referral incentives suggests longer-lasting effects of referrals. Firms might use referral incentives to encourage existing customers to refer new ones (*immediate* effect). These newly acquired customers can become better (organic) referrers than those who referred them, thus generating even more referrals in

subsequent generations (*longer-lasting* effect). Given the magnitude of the referral reinforcement effect that our studies identify, we believe that firms might greatly underestimate the return on investment of their RRPs when focusing exclusively on immediate effects (Ahrens et al. 2013; Jin and Huang 2014).

Second, referred customers should be a key priority for firms striving to improve both customer engagement and retention. These customers are key assets for firms. They show higher CLV (Van den Bulte et al. 2018), and as our results show, they also show higher referral values. Firms should therefore make sure to allocate enough resources towards these referred customers.

Third, our results offer advice to businesses that lack referrals (Leskovec et al. 2007). The generalization to incentivized settings suggests that referred customers refer more despite the inference of ulterior motives (Verlegh et al. 2013). Importantly, looking to raise customer satisfaction may not be the only path for firms to increase non-incentivized referrals. In particular, there may be gains for firms in how they *frame* encouragements to refer. Appealing to a customer's other-directed motives to refer by suggesting "you can help your friend snag a great deal with this referral!" might prove more effective than dredging self- or product-directed motives to refer ("you will help us by referring our products to your friends"). Of course, further research is needed to understand such effects fully.

2.7.2 Limitations and future research

Our research limitations open up opportunities for future study. First, our research is limited to explicit referrals and did not examine word-of-mouth forms such as customer reviews and mere mentions. Customer reviews often occur without any personal ties between customers making and receiving them, which lowers the chance that better matchmaking and social enrichment explains potential reinforcement effects. It is vital to examine the size of the

reinforcement effect as a function of the type of word-of-mouth (explicit referrals, mere mentions, customer reviews) and the relationship between giver and recipient.

Second, although we generalize the referral reinforcement effect across products and services, future research might examine the moderating effect of search, experience, and credence attributes of product and services on the magnitude of the referral reinforcement effect. Referral reinforcement effects might be stronger for search, and in particular, credence attributes where pre- and perhaps even post-consumption uncertainty of customers is high.

Third, this research focused on positive referrals and ignored negative referrals, such as warnings or encouragements to *not* buy a specific product that could arise from bad experiences, scandals or gossip. These mechanisms that influence recipients of a negative referral to further feed the negative communication chain are possibly very different from those we studied (Wetzer et al. 2007). For one, satisfaction level may play an even lesser role, especially later down the cascade when negative referrals unprompted by any personal consumption warn others. It is important to examine when and why (cascades of) negative referral reinforcement effects occur and persist.

Fourth, we investigated the size and reliability of referral reinforcement effects, but did not consider persistence *across generations* of referrals. Our data did not allow us to ascertain whether referred customers acted in the first, second or later ripple of referrals. Access to such data would allow future work to test when and how referral reinforcement effects persist, extinguish, or even grow over time.

In sum, the current research identified a sizeable and reliable referral reinforcement effect largely decoupled from the positive effects of customer satisfaction on the inclination to extend a referral. We believe that this is good news for firms aiming to improve customer engagement, and we hope that it stimulates further theory and research toward the determinants and implications of referral reinforcement effects in marketing.

Appendix of Chapter 2

Appendix 2A: Study 1 – Ridesharing.

Summary statistics data.

Table A2.1 has summary statistics data.

Table A2.1
Study 1: Summary Statistics Data (n = 200,098)

	Variable	1	2	3	4	5	6	7
1	Makes referral after intervention period (REFERRING)	-						
2	Receives referral (REFERRED)	.184	-					
3	Receives treatment (TREATED)	.022	.014	-				
4	REFERRED × TREATED	-.167	-.985	-.116	-			
5	# of past rides (PAST_RIDES)	-.005	-.004	-.016	.003	-		
6	# of weeks since last ride (RECENCY)	-.043	.055	.064	-.052	-.052	-	
7	# of weeks since account creation (TENURE)	-.043	.030	.046	-.029	.152	.563	-

Notes: Table entries are correlations: Pearson correlations between continuous variables, point-biserial (equal to Pearson) correlations between continuous and binary variables, and tetrachoric correlations between binary variables. Means and standard deviations not reported are confidential.

Code.

Mplus code for Model 2:

```
Title:
  Study 1 - Analysis of ridesharing customers

  Binary probit model

Data: FILE = "study1.dat";

Variable:
  NAMES = referring referred treated interaction past_rides recency tenure day1
     day2 day3 day4 day5 day6 hour1 hour2 hour3 hour4 hour5 hour6 hour7 hour8
     hour9 hour10 hour11 hour12 hour13 hour14 hour15 hour16 hour17 hour18 hour19
     hour20 hour21 hour22 hour23;

  categorical are referring ;  ! declares referring as categorical

Analysis:
  processors = 6 ;    ! number of processor cores/threads used
  estimator = ml;     ! ML-estimation
  link = probit ;     ! probit link function

Model:
! Model 2:  Binary probit model with satisfaction-proxies
referring ON referred treated interaction ;
referring ON past_rides recency tenure ;
referring ON day1 day2 day3 day4 day5 day6 ;
referring ON hour1 hour2 hour3 hour4 hour5 hour6 hour7 hour8 hour9 ;
referring ON hour10 hour11 hour12 hour13 hour14 hour15 hour16 hour17 ;
referring ON hour18 hour19 hour20 hour21 hour22 hour23 ;

OUTPUT:
standardized ;  ! to obtain R2 estimates
```

Appendix 2B: Study 2a – Retail banking.

Summary statistics data.

Table A2.2 contains the summary statistics data from (Ramaseshan et al. 2017) used for our reanalysis. Data on attitudinal loyalty, a construct closely related to satisfaction, were also available. Attitudinal loyalty was not included in our reanalysis because it did not express discriminant validity with satisfaction and with referring (Fornell and Larcker 1981; Pieters 2017).

Table A2.2
Study 2a: Summary Statistics Data (n = 470)

	Variable	M	SD	1	2	3
1	Receives referral (REFERRED)	.50	.50	-		
2	Customer satisfaction (SAT)	5.87	1.15	.18	.85	
3	Makes referral (REFERRING)	4.45	1.86	.28	.56	.94

Notes: Means (M), standard deviations (SD), correlations, and reliabilities are on the diagonal; "-" denotes REFERRED as a single-indicator measure; REFERRED and REFERRING are archival data from a referral-reward-program, and SAT is a survey-measure. Data are adapted from Tables 2 and 3 in Ramaseshan et al. (2017).

Model and estimation details.

We used a single-indicator structural equation model (SI-SEM) to estimate the referral reinforcement effect and the extent to which it is mediated by satisfaction. Summary statistics data (SSD) are sufficient to estimate our linear structural equation models that we use here. Our model corrects for measurement error in latent variables with composite-indicators when reliability estimates are available (Bollen 1989; Fuller and Hidiroglou 1978; MacKenzie 2001). Since reliability information was unavailable for REFERRED, we assumed it to be free of measurement error (i.e., reliability = 1). We specified a measurement model for SAT and REFERRING to account for their measurement error. The measurement model fixed the loading to one ($\lambda = 1$) and the measurement error to $(1 - \alpha)\sigma_x^2$ for identification, where α is the reliability estimate of the latent variable and σ_x^2 reflects the variance of the single-indicator x. The measurement model is:

$$x_{SAT,i} = SAT_i + \varepsilon_{SAT,i}, \text{ with } \sigma^2_{\varepsilon_{SAT}} = (1 - .85)1.15^2, \quad \text{(A2.1)}$$

$$x_{REFERRING,i} = REFERRING_i + \varepsilon_{REFERRING,i}, \text{ with } \sigma^2_{\varepsilon_{REFERRING}} = (1 - .94)1.86^2, \quad \text{(A2.2)}$$

where i indicates the customer. This model specification corrects for measurement error in the latent variables by separating the total variance of the indicators into systematic variance of the latent variable and measurement error in the σ^2 error terms. The specification is an *errors-in-variables* model (e.g., Fuller and Hidiroglou 1978) which prevents endogeneity bias due to measurement error in independent variables. The SI-SEM assumes that all observed indicators for latent variables are equally good, which seems reasonable. The structural model is then estimated simultaneously with the measurement model:

$$SAT_i = \alpha REFERRED_i + \zeta_{1,i} \quad \text{(A2.3)}$$

$$REFERRING_i = \beta SAT_i + \gamma REFERRED_i + \zeta_{2,i}, \quad \text{(A2.4)}$$

where α, β, γ are regression parameters to be estimated, and the $\zeta \sim N(0, \sigma^2_\zeta)$ terms are freely estimated structural error-terms. The model estimates the satisfaction-mediated referral reinforcement effect ($\alpha * \beta$) and non-satisfaction-mediated referral reinforcement effect (γ). The measurement model ensures that both effects are unbiased (Pieters 2017).

We estimated the 95% CIs of the mediation effects using Monte Carlo simulations with 25,000 replications (Tofighi and MacKinnon 2016). The lower and upper bounds of the 95% CI of the mediation effects are estimated using the 2.5% and 97.5% percentiles of their Monte Carlo distribution.

Code for estimation appears below. Analysis proceeded in three steps. Step 1 analyzed the summary statistics data in Mplus. Step 2 performed Monte Carlo simulations in Mplus. Step 3 uses R code to input the Mplus results since the required 95% Monte Carlo CIs are not provided by default in Mplus.

Code – Step 1.

```
TITLE:

  Study 2a

  Reanalysis of data from Ramaseshan, B., Wirtz, J., & Georgi, D. (2017).
  The Enhanced Loyalty Drivers of Customers Acquired Through Referral Reward
  Programs. Journal of Service Management, 28(4), 687-706.

  Means/standard deviations/correlations are in Table 3 of Ramaseshan et al. (2017)
  Reliabilities are in Table 2 of Ramaseshan et al. (2017)

  Variables:
   Receives referral (REFERRED)
   Customer satisfaction (SAT)
   Makes referral (REFERRING)

  Reliability of resp. SAT and REFERRING are .85 and .94

  Model is a single indicator structural equation model which fixes the
  variances of the error terms to:

  var(e.SAT) = (1-reliability(SAT))*var(SAT)
             = 1-0.85 * 1.15^2 = 0.198
  var(e.REFERRING) = (1-reliability(v3))*var(REFERRING)
             = 1-0.94 * 1.86^2 = 0.208

Data:
  FILE = study1.dat ;
  TYPE = means stdeviations correlation ; ! input is ssd
  NOBSERVATIONS = 470 ; ! sample size

Variable:
  NAMES = referred sat referring;

Analysis:
  ESTIMATOR = ML ;

Model:
  ! Measurement model
  FSAT BY sat@1 ;                    ! loadings fixed to one
  FREFERRING BY referring@1 ;
  sat@.198 ;                        ! error variances fixed as calculated above
  referring@.208 ;

  ! Structural model
  FREFERRING ON FSAT (b) ;
  FREFERRING ON referred (g) ;
  FSAT ON referred (a) ;

Model constraint:
  new(sat_mediated non_sat_mediated total_reinforcement
  perc_sat_mediated perc_non_sat_mediated) ;
  sat_mediated = a * b ;
  non_sat_mediated = g ;
  total_reinforcement = a * b + g ;
  perc_sat_mediated = sat_mediated / total_reinforcement ;
  perc_non_sat_mediated = non_sat_mediated / total_reinforcement ;

Output:
standardized; ! obtain R2 estimates
```

Code – Step 2.

```
TITLE:
Study 2a - Monte Carlo Analysis

MONTECARLO:
NAMES = referred sat referring ; ! variable names
NOBS = 470 ; ! sample size
NREPS = 25000;  ! number of replications
SEED = 1234; ! seed for replicability
GENERATE = referred (1) ; ! referred is categorical
CUTPOINTS = referred (0) ; ! cutpoint of 0
RESULTS = results.sav; ! save results in this file

ANALYSIS:
processors = 4 ; ! number of processors/threads
estimator = ML ; ! maximum likelihood estimation

MODEL POPULATION:
[referred@.5]; ! mean of referred
referred@.25;  ! variance of referred
[sat@5.87];    ! and so forth
sat@1.32;
[referring@4.45];
referring@3.46;

referred with sat@.10;  ! covariances
referred with referring@.26;
sat with referring@1.20;

MODEL:
! Measurement model
FSAT BY sat@1 ;                  ! loadings fixed to one
FREFERRING BY referring@1 ;
sat@.198 ;                       ! error variances fixed
referring@.208 ;

! Structural model
FREFERRING ON FSAT (b) ;
FREFERRING ON referred (g) ;
FSAT ON referred (a) ;

MODEL CONSTRAINT:
new(sat_mediated non_sat_mediated total_reinforcement
perc_sat_mediated perc_non_sat_mediated) ;
sat_mediated = a * b ;
non_sat_mediated = g ;
total_reinforcement = a * b + g ;
perc_sat_mediated = sat_mediated / total_reinforcement ;
perc_non_sat_mediated = non_sat_mediated / total_reinforcement ;

OUTPUT: tech1; ! tech1 for parameter labels;
```

Code – Step 3.

```
rm(list=ls(all=TRUE)) # clear workspace
# change working directory
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
options(scipen=999) # disable scientific notation

source = readLines("results.sav") # read the Mplus results
nreps = 25000 # number of replications
results = list() # define a list for the results

for(i in 1:nreps){ # start loop over number of replications
start = 1+(i-1)*7  # output for each replication has 7 lines
end = start+6
represults <- as.numeric(unlist(strsplit( # read the estimates
                       source[start:end], " ")))
results[[i]] = represults # enter them in the list
} # end loop

results = do.call(rbind, results) # list to dataframe

# Satisfaction-mediated effect
mean(results[,9]) # mean across replications
quantile(results[,9], c(.05/2, 1-.05/2)) # 95% MC CI

# Non-satisfaction-mediated effect
mean(results[,10])
quantile(results[,10], c(.05/2, 1-.05/2))

# Total effect
mean(results[,11])
quantile(results[,11], c(.05/2, 1-.05/2))
```

Appendix 2C: Study 2b – Movies.

Measurement details.

Table A2.3 lists details on the items and scales used in Study 2b.

Table A2.3
Study 2b: Questionnaire Items and Response Scales

Construct	Item(s)	Response scale
Receives referral (REFERRED)	Did anyone recommend you this movie before you saw it?	No (0) / Yes (1)
Makes referral (REFERRING)	Did you recommend the movie to anyone after seeing it?	No (0) / Yes (1)
	Do you intend to recommend the movie to anyone in the future?	No (0) / Yes (1)
Movie satisfaction (SAT)	I am satisfied with my overall experience with the movie	Strongly Disagree (1) - Strongly Agree (7)
	As a whole, I am not satisfied with the movie [a]	Strongly Disagree (7) - Strongly Agree (1)
	How satisfied are you overall with the quality of the movie?	Very Dissatisfied (1) - Very Satisfied (7)
Opinion seeking (SEEK)	When I consider seeing a movie, I ask other people for advice	Strongly Disagree (1) - Strongly Agree (7)
	I don't need to talk to others before I see a movie [a]	Strongly Disagree (7) - Strongly Agree (1)
	I rarely ask other people what movies to see [a]	Strongly Disagree (7) - Strongly Agree (1)
	I like to get others' opinions before I see a movie	Strongly Disagree (1) - Strongly Agree (7)
	I feel more comfortable seeing a movie when I have gotten other people's opinions on it	Strongly Disagree (1) - Strongly Agree (7)
	When choosing a movie, other people's opinions are not important to me [a]	Strongly Disagree (1) - Strongly Agree (7)
Opinion leadership (LEADER)	My opinion about movies seems not to count with other people [a]	Strongly Disagree (7) - Strongly Agree (1)
	When they choose a movie, other people do not turn to me for advice [a]	Strongly Disagree (7) - Strongly Agree (1)
	Other people rarely come to me for advice about choosing movies [a]	Strongly Disagree (7) - Strongly Agree (1)
	People that I know pick movies based on what I have told them	Strongly Disagree (1) - Strongly Agree (7)
	I often persuade other people to see movies that I like	Strongly Disagree (1) - Strongly Agree (7)
	I often influence people's opinions about popular movies	Strongly Disagree (1) - Strongly Agree (7)
Gender (GENDER)	What is your gender?	Female (0) / Male (1)
Age (AGE)	What is your age?	18-65

[a] : negatively worded (reversed) item

Measurement model.

Prior to estimating the structural model, a confirmatory factor analysis (CFA) validated the multi-item scales. The CFA had a satisfactory fit ($\chi^2_{(87)}$ = 947; CFI = .88; RMSEA = .11; SRMR = .06). Composite reliabilities were .77, .88, .87 for SAT, SEEK and LEADER, respectively. The average variance extracted (AVE) was .53, .55, .54, respectively. AVEs significantly below the variance shared for all combinations of factors established discriminant validity (Fornell and Larcker 1981; Pieters 2017).

Summary statistics data.

Table A2.4 has summary statistics data.

Table A2.4
Study 2b: Summary Statistics Data (n = 851)

	Variable	M	SD	1	2	3	4	5	6	7	8	9
1	Makes referral (REFERRING)	.732	.443	-								
2	Receives referral (REFERRED)	.376	.485	.384	-							
3	Movie satisfaction (SAT)	5.724	1.203	.516	.157	.77						
4	Opinion seeking (SEEK)	3.848	7.134	.141	.241	.015	.88					
5	Opinion leadership (LEADER)	4.380	5.757	.297	.143	.149	.367	.87				
6	Opening weekend box-office revenue (BOX)	17.508	1.464	.022	.140	.016	.099	-.008	-			
7	Movie rating (RATE)	75.566	9.994	.273	.307	.291	.077	.033	.190	-		
8	Gender (GENDER)	.402	.491	.052	.003	.043	-.022	-.044	-.069	-.087	-	
9	Age (AGE)	31.973	9.470	.065	-.092	.069	-.083	-.020	-.087	-.115	.121	-

Notes: Table entries present means (M), standard deviations (SD), and correlations with estimated reliabilities on the diagonal. "-" denotes a reliability estimate not available. Reported correlations list Pearson correlations between continuous variables, point-biserial (equal to Pearson) correlations between continuous and binary variables, and tetrachoric correlations between binary variables. BOX is the natural logarithm of a movie's opening weekend box-office revenues in U.S. dollars.

Estimation details.

We conducted Bayesian mediation analysis (Zhang et al. 2009) to accommodate the fact that the mediation model includes three latent variables (LEADER, SEEK and SAT), each measured with multiple items, and two categorical dependent variables (REFERRED and REFERIN), which prevented standard ML estimation. Bayesian mediation analysis provides mean estimates similar to frequentist models that rely on bootstrapping, but can handle more complex models and provide precise estimates even at comparatively small sample sizes. We enlisted Bayesian estimation (with 25,000 MCMC iterations and default non-informative priors) in the Mplus software version 8.3 (Muthén and Muthén 2018) to estimate parameters. To properly estimate the indirect and direct effects on the categorical dependent variable, we back-transformed Probit estimates as specified in Muthén et al. (2016).

Code.

The Mplus code to estimate the full Bayesian mediation model (Model 2):

```
Title:
  Study 2b - Analysis of 851 moviegoers
  Generalized structural equation model – Bayesian estimation
```

```
Data:
  File is study2b.dat ;

Variable:
  Names are                          ! variable labels are self-explanatory
     satisfaction1 satisfaction2 satisfaction3 seeking1 seeking2 seeking3
     seeking4 seeking5 seeking6 leader1 leader2 leader3 leader4 leader5 leader6
     gender age rate box referred referin ;

  categorical are referred referin ;  ! declares categorical

Analysis:
  processors = 6 ;    ! number of processor cores/threads used
  estimator = bayes;  ! note: default link in Bayes Mplus for categorical is probit
  point = mean;       ! default is median, change to mean
  fbiter = (25000);   ! 25,000 MCMC draws

Model:
  ! Measurement model
  SAT by satisfaction1-satisfaction3*; SEEK by seeking1-seeking6*;
  LEADER BY leader1-leader6*;

  ! Measurement model means; fixed to zero for identification
  [satisfaction1-satisfaction3@0]; [seeking1-seeking6@0]; [leader1-leader6@0];

  ! Latent variances (fixed to one for identification)
  SAT@1; SEEK@1; LEADER@1;

  ! Latent means
  [SEEK*] (meanseek); [LEADER*] (meanlead);

  ! Manifest means
  [box*] (meanbox); [rate*] (meanrate); [gender*] (meangender); [age*] (meanage);

  ! Equation 2.2: referred
  referred ON SEEK LEADER box rate gender age ;

  ! Equation 2.3: SAT
  SAT ON SEEK (omega21); SAT ON LEADER (omega22); SAT ON box (omega23);
  SAT ON rate (omega24); SAT ON gender (omega25); SAT ON age (omega26);
  SAT ON referred (alpha) ;
  [SAT*] (omega20); ! intercept of SAT

  ! Equation 2.4: referring
  referin ON SEEK (omega31); referin ON LEADER (omega32); referin ON box (omega33);
  referin ON rate (omega34); referin ON gender (omega35); referin ON age (omega36);
  referin ON SAT (beta); referin ON referred (gamma);
  [referin$1*] (omega30) ;        ! threshold (-intercept) of referring

Model constraint:
  NEW(cov2 cov3 arg11 arg10 arg00 satmed nonsatmed total percm percnonm) ;

!  For details of the total effect decomposition see Chapter 8 in:
!  Muthén, B. O., Muthén, L. K., & Asparouhov, T. (2016).
!    Regression and Mediation Analysis Using Mplus (1st ed.).
!    Los Angeles, CA: Muthén & Muthén.

! Linear indices for covariates
  cov2 = omega21 * meanseek + omega22 * meanlead + omega23 * meanbox +
         omega24 * meanrate + omega25 * meangender + omega26 * meanage ;

  cov3 = omega31 * meanseek + omega32 * meanlead + omega33 * meanbox +
         omega34 * meanrate + omega35 * meangender + omega36 * meanage ;

! Expressions for the counterfactually defined linear indices
  arg11 = -omega30 + gamma*1 + cov3 + beta*(omega20 + alpha*1 + cov2);
  arg10 = -omega30 + gamma*1 + cov3 + beta*(omega20 + alpha*0 + cov2);
  arg00 = -omega30 + gamma*0 + cov3 + beta*(omega20 + alpha*0 + cov2);
```

```
! Total effect decomposition (phi is the standard normal distribution function)
  ! Satisfaction-mediated
  satmed = phi(arg11/sqrt(beta^2 * 1 + 1)) - phi(arg10/sqrt(beta^2 * 1 + 1));
  ! Non-satisfaction-mediated
  nonsatmed = phi(arg10/sqrt(beta^2 * 1 + 1)) - phi(arg00/sqrt(beta^2 * 1 + 1));
  ! Total effect
  total = phi(arg11/sqrt(beta^2 * 1 + 1)) - phi(arg00/sqrt(beta^2 * 1 + 1));
  ! Percentage satisfaction-mediated
  percm = (satmed / total) * 100 ;
  ! Percentage non-satisfaction-mediated
  percnonm = (nonsatmed / total) * 100 ;

OUTPUT:
standardized ; ! to obtain R-square estimates
```

Detailed estimation results.

Table A2.5 details the estimation results for the models with (Model 1) and without (Model 2) covariates.

Robustness checks and alternative explanations.

First, we extended Model 2 (with covariates) by estimating the interaction effect between satisfaction and referral reception on likelihood of REFERRING (Equation 2.4) using a latent interaction specification (Asparouhov and Muthén 2020). The interaction between satisfaction and referral reception did not significantly affect the likelihood to refer others (β = -.15, one-tailed p = .17) beyond the two main effects. This rules out that satisfied customers who had been referred might have a stronger inclination to refer others; if anything, it was weaker.

Second, over and above the linear effect, the effect of a squared satisfaction term did not significantly influence the inclination to refer (β = .04, p = .20). This rules out the possibility that in particular customers with very high (low) satisfaction levels are more (less) likely to refer (Anderson 1998).

Third, replacing the opening weekend box-office revenue with the log of total box-office revenue did not change results ($\Phi(\alpha*\beta)$ = .03, 95% CI [.01, .06]; $\Phi(\gamma)$ = .05, 95% CI [.01, .08]). This rules out findings shaped by the revenue period.

Fourth, a follow-up analysis with only referrals from stronger ties (one's partner, family members and/or a friend, n = 813 out of 851) gave very similar findings as our main analysis did ($\Phi(\alpha*\beta)$ = .03, 95% CI [.01, .05]; $\Phi(\gamma)$ = .07, 95% CI = [.02, .09]). It rules out that the referral reinforcement effect is only due to referrals of weaker ties.

Table A2.5
Study 2b: Referral Reinforcement Effects among Moviegoers (n = 851)

		Model 1			Model 2		
Variable	Parameter	Estimate	SD	*P*-value	Estimate	SD	*P*-value
Receives referral (REFERRED)							
Intercept	$\omega_{1,0}$				-4.957	(.660)	<.001
Opinion seeking (SEEK)	$\omega_{1,1}$				.299	(.059)	<.001
Opinion leadership (LEADER)	$\omega_{1,2}$				.064	(.057)	.132
Opening weekend box-office revenue (BOX)	$\omega_{1,3}$				.033	(.031)	.138
Movie rating (RATE)	$\omega_{1,4}$				.039	(.005)	<.001
Gender (GENDER)	$\omega_{1,5}$				.105	(.096)	.137
Age (AGE)	$\omega_{1,6}$				-.008	(.005)	.058
Movie satisfaction (SAT)							
Intercept	$\omega_{2,0}$	5.324	(.151)	<.001	2.087	(.547)	<.001
Opinion seeking (SEEK)	$\omega_{2,1}$				-.109	(.048)	.011
Opinion leadership (LEADER)	$\omega_{2,2}$				.199	(.045)	<.001
Opening weekend box-office revenue (BOX)	$\omega_{2,3}$				-.022	(.026)	.193
Movie rating (RATE)	$\omega_{2,4}$				.034	(.004)	<.001
Gender (GENDER)	$\omega_{2,5}$				.202	(.076)	.004
Age (AGE)	$\omega_{2,6}$				.014	(.004)	<.001
Receives referral (REFERRED)	α	.423	(.075)	<.001	.151	(.053)	.002
Makes referral (REFERRING)							
Intercept	$\omega_{3,0}$	-4.465	(.411)	<.001	-7.309	(.824)	<.001
Opinion seeking (SEEK)	$\omega_{3,1}$				.013	(.080)	.434
Opinion leadership (LEADER)	$\omega_{3,2}$				.471	(.078)	<.001
Opening weekend box-office revenue (BOX)	$\omega_{3,3}$				-.055	(.040)	.084
Movie rating (RATE)	$\omega_{3,4}$				.013	(.007)	.032
Gender (GENDER)	$\omega_{3,5}$				-.013	(.123)	.455
Age (AGE)	$\omega_{3,6}$				.012	(.007)	.039
Movie satisfaction (SAT)	β	.922	(.079)	<.001	.818	(.083)	<.001
Receives referral (REFERRED)	γ	.440	(.120)	<.001	.187	(.083)	.011

Referral reinforcement effect decomposition	Parameter	Estimate	95% CI	Estimate	95% CI
SAT-mediated reinforcement effect	$\Phi(\alpha*\beta)$	.083	[.053, .116]	.035	[.011, .058]
Non-SAT-mediated reinforcement effect	$\Phi(\gamma)$	.114	[.054, .174]	.050	[.008, .087]
Total reinforcement effect	$\Phi(\alpha*\beta+\gamma)$	.197	[.137, .255]	.085	[.039, .122]
% SAT-mediated reinforcement effect	$\Phi((\alpha*\beta)/(\alpha*\beta+\gamma))$	43%		43%	
% non-SAT-mediated reinforcement effect	$\Phi(\gamma/(\alpha*\beta+\gamma))$	57%		57%	

Notes: Latent variables identified by fixing variance to unity. Estimates of regression and path weights are unstandardized with posterior standard deviations (SD) and one-tailed Bayesian *p*-values. The referral reinforcement effect decomposition yielded estimates and 95% CIs. $\Phi(\cdot)$ denotes effects on the latent response variable back-transformed along Probit probability curve. R^2 estimates for Model 1 were .042 for SAT and .505 for REFERRING. R^2 estimates for Model 2 were .241, .191, and .606 for REFERRED, SAT and REFERRING, respectively.

Appendix 2D: Study 3 – Commercials

Measurement model.

Prior to estimating the structural model, a confirmatory factor analysis ($\chi^2_{(87)}$ = 164; CFI = .92; RMSEA = .10, SRMR = .06) showed good composite reliabilities (.94, .80, .84) and AVEs (.68, .51, .57) for AFF, COG, and REFERRING. The lower fit of models that fixed inter-factor correlations to one, in comparison to the predicted three-factor model, established discriminant validity (Pieters 2017).

Summary statistics data.

Table A2.6 has summary statistics data.

Table A2.6
Study 3: Summary Statistics Data (n = 87)

	Variable	M	SD	1	2	3	4
1	Makes referral (REFERRING)	6.954	2.028	.84			
2	Receives referral (REFERRED)	.494	.503	.242	-		
3	Affective evaluation (AFF)	8.240	1.791	.682	.088	.94	
4	Cognitive evaluation (COG)	7.787	1.666	.647	.061	.622	.80

Notes: Table entries present means (M), standard deviations (SD), and correlations with estimated reliabilities on the diagonal. Reported correlations are Pearson correlations between continuous variables and point-biserial (equal to Pearson) correlations between continuous and binary variables. "-" denotes an estimate for the REFERRED manipulation not available.

Code.

The Mplus code to estimate the structural equation model with raw data:

```
Title:
  Study 3 - Analysis of 87 participants in the lab

Data:
  File is study3.dat ;

Variable:
   Names are  ! variable labels are self-explanatory
     referred referring1 referring2 referring3 referring4
     cog1 cog2 cog3 cog4
     aff1 aff2 aff3 aff4 aff5 aff6 aff7 ;

Analysis:
  estimator = ml ;     ! maximum-likelihood estimation
  Processors = 6 ;     ! 6 parallel processors / threads
  bootstrap = 25000 ;  ! 25,000 bootstrap replications

Model:
! Measurement model
  COG by cog1* cog2 cog3 cog4 ;
  AFF by aff1* aff2 aff3 aff4 aff5 aff6 aff7 ;
  REFERRING by referring1* referring2 referring3 referring4 ;
```

```
! (Co)variances
  COG@1; AFF@1; REFERRING@1; ! variances to 1 for identification
  COG WITH AFF* (cov);         ! free error covariance between mediators

! AFF
  AFF ON referred (a1) ;

! COG
  COG ON referred (a2) ;

! REFERRING
  REFERRING ON AFF (b1) ;
  REFERRING ON COG (b2) ;
  REFERRING ON referred (g) ;

Model constraint:                  ! Total effect decomposition
  new(sat_mediated1 sat_mediated2
  non_sat_mediated total_reinforcement
  perc_sat_mediated perc_non_sat_mediated) ;
  sat_mediated1 = a1 * b1 ;
  sat_mediated2 = a2 * b2 ;
  non_sat_mediated = g ;
  total_reinforcement = a1 * b1 + a2 * b2 + g ;
  perc_sat_mediated = sat_mediated1 + sat_mediated2 / total_reinforcement ;
  perc_non_sat_mediated = non_sat_mediated / total_reinforcement ;

Output:
cinterval(bcbootstrap);  ! bootstrap CIs
standardized ; ! R2 estimates
```

Appendix 2E: Study 4 – Customer lay beliefs

Scenario.

For example, the referred and extremely satisfied scenario was scripted as:

"Imagine that a new product has just been introduced. The product is made by a manufacturer that you do not know. Not many people have tried the product yet, but you are considering to buy the product.

You receive a recommendation for the product from a friend who knows you well. This person has bought the product already and tells you that you should definitely get it too. This person told you: "You should really buy this product, this is something you cannot miss!"

You decide to follow the recommendation and buy the product.

You really like the product. You are exceptionally happy with it. You are feeling extremely satisfied with the product.

The next day, after buying the product, you are meeting with another friend who does not have the product yet. You are thinking about the product you bought. You think about the product and the experience you had. At this point in time, you decide to give your friend a recommendation for the product: "I bought a product that just came out, you should get it as well." You recommend this friend to buy the product. "

Measurement model details.

Table A2.7 lists the full set of 23 items, factor loadings, and summary statistics.

Table A2.7
Study 4: Referral Motive Items (n = 1,210)

Item	Factor	Standardized factor loading	M	SD
You will get to know your friend better.	Self-directed	.821	2.492	1.183
It will strengthen your relationship with your friend.	Self-directed	.805	2.628	1.239
Your friend will get to know you better.	Self-directed	.804	2.481	1.155
You want to have a nice chat with your friend.	Self-directed	.777	2.973	1.267
You want to keep the conversation with your friend going.	Self-directed	.765	2.760	1.254
You will have an experience in common with your friend.	Self-directed	.733	3.155	1.255
It stimulates that your friend will give you product recommendations in return.	Self-directed	.714	2.483	1.198
It makes you look good.	Self-directed	.611	2.305	1.202
Your friend's experience with this product would amplify your own experience with it.	Self-directed	.575	2.883	1.223
You will learn from your friend's experience with this product.	Self-directed	.540	2.983	1.255
You received an incentive (e.g., a discount, gift, special service) to recommend this product.	Self-directed	.211	2.155	1.266
You want to do something nice for your friend.	Other-directed	.786	3.102	1.295
It will help your friend deciding whether to buy this product or not.	Other-directed	.774	3.367	1.258
It feels right to do so.	Other-directed	.724	2.996	1.270
Your friend will like this product.	Other-directed	.656	3.698	1.172
Your friend knows others who will like this product.	Other-directed	.651	2.902	1.199
You want to make sure that your friend will buy this product.	Other-directed	.559	2.335	1.157
Your friend asked you to recommend a product.	Other-directed	.474	3.013	1.319
You feel obligated to do so.	Other-directed	.434	2.162	1.133
You enjoy talking about this product.	Product-directed	.749	2.864	1.279
You want to relive the experience of using this product yourself.	Product-directed	.736	2.240	1.166
You will help this product to become successful.	Product-directed	.712	2.293	1.205
You feel that you cannot keep your experience with this product to yourself.	Product-directed	.664	2.830	1.264

Notes: Standardized factor loadings from a three-factor confirmatory factor analysis (3-CFA). All loadings of items on their respective factors were statistically significant at $p < .001$. Latent variables identified by fixing variance to unity. Composite reliabilities were .90, .85, and .81, respectively, for self-, other- and product-directed motives. Latent correlations were .78 for self- with other-directed, .86 for self- with product-directed, and .83 for product- with other-directed motives. M refers to the mean and SD to the standard deviation of each item.

Code.

The Mplus code to estimate the latent MANOVA:

```
TITLE: Study 4 - Analysis of Lay Beliefs (n = 1,210)

DATA: FILE = "study4.dat";

VARIABLE:
NAMES = referred sat interaction self1 self2 self3 self4
        self5 self6 self7 self8 self9 self10 self11
        other1 other2 other3 other4 other5 other6
        other7 other8 product1 product2 product3 product4;

ANALYSIS:
estimator = ML; ! maximum-likelihood estimation

MODEL:
! Measurement model
SELF by self1-self11*;
OTHER by other1-other8*;
PRODUCT by product1-product4*;

SELF@1; OTHER@1; PRODUCT@1; ! variances fixed to 1 to identify

! Structural model: Latent MANOVA
SELF OTHER PRODUCT on referred sat interaction ;

OUTPUT:
standardized; ! obtain R2 estimates
```

Estimation results.

Table A2.8 presents the estimates from the latent MANOVA.

Table A2.8
Study 4: Latent MANOVA on Lay Beliefs about Motives to Refer When (Not) Having Been Refered Oneself (n = 1,210)

Variable	Parameter	Estimate	SE	*P*-value
Self-directed motives				
Receives referral (REFERRED)	$\gamma_{1,1}$	.034	(.030)	.260
Customer satisfaction (SAT)	$\beta_{1,1}$	.246	(.022)	<.001
REFERRED × SAT	$\beta_{1,2}$	-.031	(.021)	.138
Other-directed motives				
Receives referral (REFERRED)	$\gamma_{2,1}$	.134	(.032)	<.001
Customer satisfaction (SAT)	$\beta_{2,1}$	.447	(.025)	< 001
REFERRED × SAT	$\beta_{2,2}$	-.061	(.022)	.006
Product-directed motives				
Receives referral (REFERRED)	$\gamma_{3,1}$	-.008	(.033)	.816
Customer satisfaction (SAT)	$\beta_{3,1}$	.498	(.028)	<.001
REFERRED × SAT	$\beta_{3,2}$	-.042	(.023)	.073

Notes: Table entries list unstandardized parameter estimates, standard errors (SE) and two-tailed *p*-values from a latent MANOVA. Latent variables identified by fixing the variance to unity. R^2 estimates were .113, .305 and .338, respectively, for self-, other-, and product-directed motives.

Post-hoc analyses of the significant interaction effect of REFERRED × SAT on other-directed motives revealed that the referred condition differed from the non-referred condition only for the extremely dissatisfied case (M_No = 2.47 (.86), M_Yes = 2.81 (.83); difference $p < .01$) and dissatisfied case (M_No = 2.69 (.82), M_Yes = 2.89 (.78); difference $p = .03$) as seen in Table 2.6, but not for the other satisfaction conditions (all $p > .48$).

Appendix 2F: Meta-effect estimation and forest plot.

We calculated the meta-effect as follows. First, we corrected estimated correlations between REFERRED and REFERRING for measurement error in Studies 1, 2a, 2b and 3. Second, we transformed the correlations to Fisher-Z values and calculated the standard error and 95% CI. Third, we took the simple and weighted (by the standard error) means of Fisher-Z values and back-transformed them to meta-analytic correlations. We present both simple and weighted means to account for large differences in sample sizes among the studies (min = 87, max = 200,098). Finally, Step 4 plots the results. The R code:

```
# R code to calculate the meta-effect and make a forest plot

rm(list=ls(all=TRUE)) # clear workspace
# change working directory
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
options(scipen=999) # disable scientific notation

# Step 1: Enter input
# Correlations (Study 1, S2a, S2b, S3) - without accounting for covariates
r = c(.184, .28/sqrt(1*.94), .384, .242/sqrt(1*.84))
# Sample sizes
n = c(200098, 470, 851, 87)

# Step 2: Fisher-Z
# Fisher-Z transformation
Zr = atanh(r)
# Standard deviation of Z
SDZr = sqrt(n-3)
# 95% CI of Z
upperZr = Zr + qnorm(1-.05/2)*1/SDZr
lowerZr = Zr - qnorm(1-.05/2)*1/SDZr
# Back-transformed 95% CI of Z
upperr = atan(upperZr)
lowerr = atan(lowerZr)

# Step 3: Mean Z
# Average Z (uses simple mean)
MZr = mean(Zr)
# Standard deviation of average Z
SDMZr = sqrt(sum(n)-3)
# 95% CI of Z
upperMZr = MZr + qnorm(1-.05/2)*1/SDMZr
lowerMZr = MZr - qnorm(1-.05/2)*1/SDMZr
# Back-transformed Z
Mr = atan(MZr)
# Back-transformed 95% CI of Z
upperMr = atan(upperMZr)
lowerMr = atan(lowerMZr)

# Average Z (uses weighted, by SD, mean)
wMZr = sum(Zr*sqrt(n-3)) / sum(sqrt(n-3))
# Standard deviation of average Z
SDwMZr = sqrt(sum(n)-3)
# 95% CI of Z
upperwMZr = wMZr + qnorm(1-.05/2)*1/SDwMZr
lowerwMZr = wMZr - qnorm(1-.05/2)*1/SDwMZr
```

```
# Back-transformed Z
wMr = atan(wMZr)
# Back-transformed 95% CI of Z
upperwMr = atan(upperwMZr)
lowerwMr = atan(lowerwMZr)

# Step 4: Plot
# Prepare plot data
names = factor(x = 1:6, levels = rev(1:6),
          labels = rev(c("Study 1: Ridesharing\nn = 200,098",
                         "Study 2a: Retail banking\nn = 470",
                         "Study 2b: Movies\nn = 851",
                         "Study 3: Commercials\nn = 87",
                         "Meta-effect (simple mean)\nn = 201,506",
                         "Meta-effect (weighted mean)\nn = 201,506")),
          ordered = T)
lower = c(lowerr, lowerMr, lowerwMr)
est = c(r, Mr, wMr)
upper= c(upperr, upperMr, upperwMr)
label = sapply(est, function(x) sub("^0+","",formatC(x,format='f',digits=2)))
plotdata = data.frame(names, lower, est, upper) # data frame holds all plotdata

# Start plotting using library "ggplot2"
library(ggplot2)
ggplot(data = plotdata, aes(y = names, x = est,
                            xmin = lower, xmax = upper)) +
  geom_vline(aes(xintercept = 0), linetype = "dashed", size = .25) +
  geom_errorbarh(height = .25, size = .25) +
  geom_point(size = .75) +
  xlab("Effect size (r)") +
  ylab("Referral Reinforcement Effect") +
  theme_bw() +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  scale_x_continuous(breaks = c(0, .1, .2, .3, .4, .5),
                     labels = c("0", ".10", ".20", ".30", ".40", ".50")) +
  coord_cartesian(xlim = c(0,.50)) +
  geom_text(aes(label = label), vjust = c(2.5,2,2,2,2.5,2.5), size = 2.5)
ggsave("forestplot.png", width = 15, height = 10, units = "cm") # save the plot
```

Chapter 3 – Six Moderation Analysis Methods for Marketing Research: A Comparison

3.1 Introduction

Managers and researchers often want to know the effect of a decision variable X on a performance variable Y and whether this effect depends on a moderator Z. Moderation analysis promises to provide important insights into the boundary conditions of relationships between managerially relevant variables and offer deeper process insights (Goldsby et al. 2013). Decision variables and moderators, which frequently are latent and unobserved by analysts, tend to be measured with one or more indicators that contain random measurement error. For example, the effect of brand differentiation as a latent variable on profits is moderated by market uncertainty, another latent variable (Dahlquist and Griffith 2014). Similarly, the effect of brand extension fit on brand extension success depends on the quality of the parent brand (Völckner and Sattler 2006), and the influence of affective responses to an ad on persuasion by the ad changes with regulatory focus (Haws et al. 2010). In each of these examples, the studies measure the latent interacting variables X and Z using multiple indicators with measurement error.

Valid theory testing and policy planning and evaluation require that estimated moderation effects be unbiased, with accurate inferences about their size, sign, and statistical significance. That is, we need to estimate moderation effects with small estimation bias and large statistical power. Estimation bias reflects the discrepancy between the estimated and true moderation effect, so a smaller bias is relevant for quantifying the theoretical and managerial impact of the moderating variable. Strategic decisions based on biased estimates of moderation effects can fail to deliver the expected results or are inefficient. Maximizing statistical power, the probability that a true moderation effect is estimated as statistically significant, is also important to identify a true moderation effect. Failing to identify a

significant moderation effect or its size can be particularly damaging when the goal is to generalize an effect or its boundary conditions.

Several methods are available to test for moderation effects in the presence of measurement error (Klein and Moosbrugger 2000; Marsh et al. 2004; Ping 1995). The methods vary in their assumptions and approaches to error in the measured indicators of the latent variables. In fact, as Grewal et al. (2004, p. 528) point out, "[e]ven when reliability is fairly high by conventional standards, measurement error can be damaging." Specifically, random measurement error in explanatory variables can induce bias due to endogeneity, increase the standard errors of the estimates, and reduce statistical power (Blalock 1965; Bollen 1989; Busemeyer and Jones 1983; Cole and Preacher 2014; Greene 2008; Grewal et al. 2004; Spearman 1904).[2] Despite this, a comprehensive assessment of the performance of the various moderation analysis methods for moderation analysis in terms of bias and power of the estimated moderation effects is as of yet unavailable. This provides managers and researchers little guidance in their choice of method—which would be particularly worrisome if the various methods perform differently in terms of bias and power.

To close this knowledge gap, we compare six common moderation analysis methods with respect to their bias and power in the presence of measurement error, and then provide recommendations for their use. Two common methods do not adequately account for measurement error (means and multi-group), and four methods do, in different ways (factor scores, corrected means, product indicators, and latent product). From a literature review, we determine which methods are most popular in marketing research, according to their use in the 504 moderation effects published in *Journal of Marketing* and *Journal of Marketing Research* between 2000 and 2017. We find that 89% of the moderation effects were tested

[2] To illustrate endogeneity due to measurement error, suppose that of interest is the model $y = x_t\beta + u$, but the observed x is only observable with random measurement error: $x = x_t + \varepsilon$. Then the model becomes: $y = x\beta + (u - \varepsilon\beta)$. Here, ε becomes part of the error term and endogeneity is due to the correlation between x and ε.

with means or multi-group methods, suggesting they are likely to be biased, underpowered, or both. The remaining 11% of published moderation analyses account for measurement error with one of the four other methods.

We also used Monte Carlo simulations to assess the bias and power of the methods in realistic conditions, with the results from the literature review as input. Even when the reliability of the interacting variables reaches .80, the means and multi-group methods provide estimates that are biased downward by more than 30%. Therefore, we generally recommend against using these two methods. All four methods that attempt to account for measurement error overall offer minimal bias. The latent product method, which is computationally intensive, achieved the lowest bias (about 1%) and highest statistical power. According to our simulations, it requires 254 observations to find a moderation effect of .20 with sufficient statistical power (80%); this sample size is about 50% larger than the median sample size of 171 in our literature review. Thus, we recommend tests of moderation with larger samples, to obtain adequate statistical power. The corrected means and product indicators methods achieve lower power than the latent product method and are unbiased in large samples (e.g., 1,500 observations), but they exhibit substantial bias with smaller samples (e.g., 175 observations). The factor scores method, which is accessible, performs remarkably well, with similar bias (about 1%) and only slightly lower power than the latent product method. It is therefore a reasonable substitute that we strongly recommend. Surprisingly, these analyses also reveal that multicollinearity between the interacting variables X and Z increases rather than decreases the statistical power of the moderation effect. The combined findings of our literature review and Monte Carlo simulations thus enable us to contribute new insights and recommendations for researchers.

3.2 Moderation Analysis in the Face of Measurement Error

Suppose that a variable Z is hypothesized to moderate the effect of a variable X on a dependent variable Y, and let XZ represent the interaction of X with Z. Without loss of generality, we assume that Y is an observed variable, with a single indicator. For example, Seiders et al. (2005) find that consumer involvement (Z: 3 indicators, reliability = .89) moderates the effect of satisfaction (X: 3 indicators, reliability = .90) on repurchase spending (Y), measured with a single indicator. The true scores of X and Z are not directly observed by the analyst, but each variable is measured with three continuous indicators, v_{X1}–v_{X3} and v_{Z1}–v_{Z3}. For this study, we use three indicators for both X and Z, as is common (Peterson 1994). Appendix 3A contains hypothetical data to test for moderation in this case. For example, Y could be repurchase spending, v_{X1}–v_{X3} could be three indicators of satisfaction (X), and v_{Z1}–v_{Z3} could be three indicators of involvement (Z). We present a framework and six moderation analysis methods for this common situation.

3.2.1 Framework of moderation.

Consider the structural regression model:

$$Y^g = \beta_1^g X^g + \beta_2^g Z^g + \beta_3^g XZ^g + \zeta^g, \tag{3.1}$$

where the subscript i for the unit of analysis is dropped for brevity; β_1^g and β_2^g are the main effects of X and Z on Y, and β_3^g is the moderation effect of X and Z on Y; and ζ^g is the residual. The superscript g ϵ (1, 2, …, G) denotes group membership (e.g., country), across which variables and parameters can vary (see Method 1.2 subsequently).

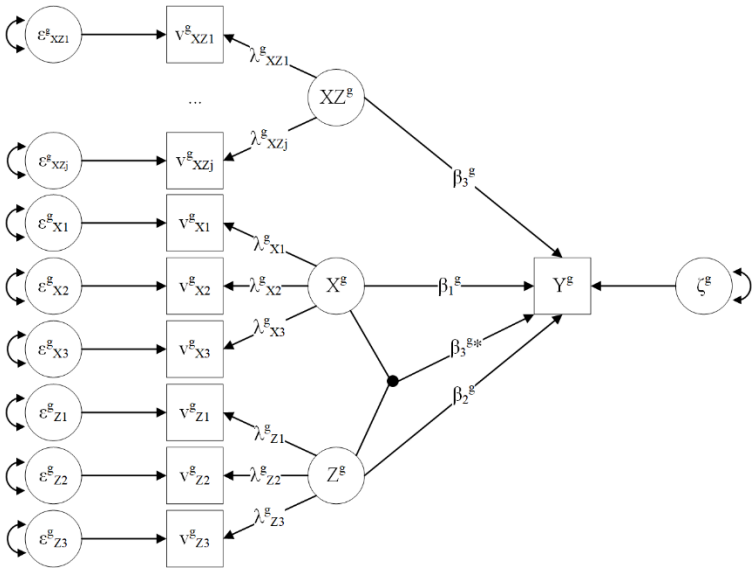
The true scores in X^g, Z^g, and XZ^g are inferred from multiple indicators, thus for X:

$$v_X^g = \lambda_X^g X^g + \varepsilon_X^g, \tag{3.2}$$

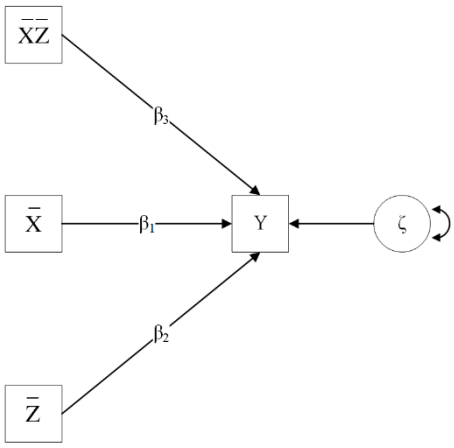
where the subscript j for the indicator is dropped for brevity, v is the indicator, λ is a factor loading, and ε is a random independent measurement error, distributed as $N(0, \sigma_\varepsilon^{2,g})$.

Figure 3.1
Graphical Representation of Six Methods for Moderation Analysis

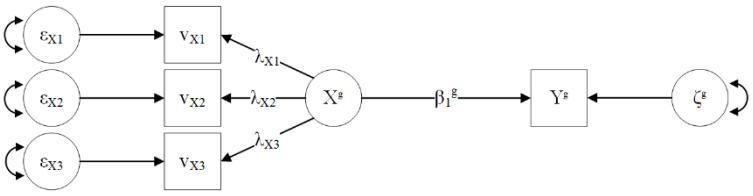
Panel A: Framework



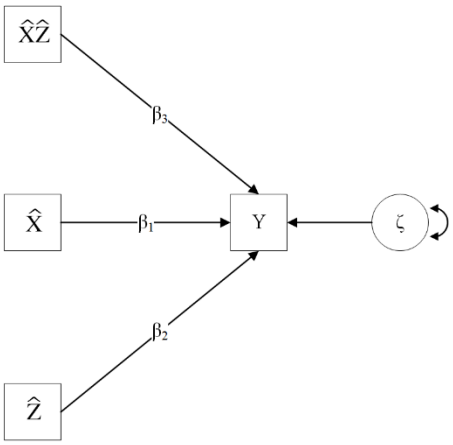
Panel B: 1.1 Means



Panel C: 1.2 Multi-group



Panel D: 2.1 Factor scores



Panel E: 2.2 Corrected means

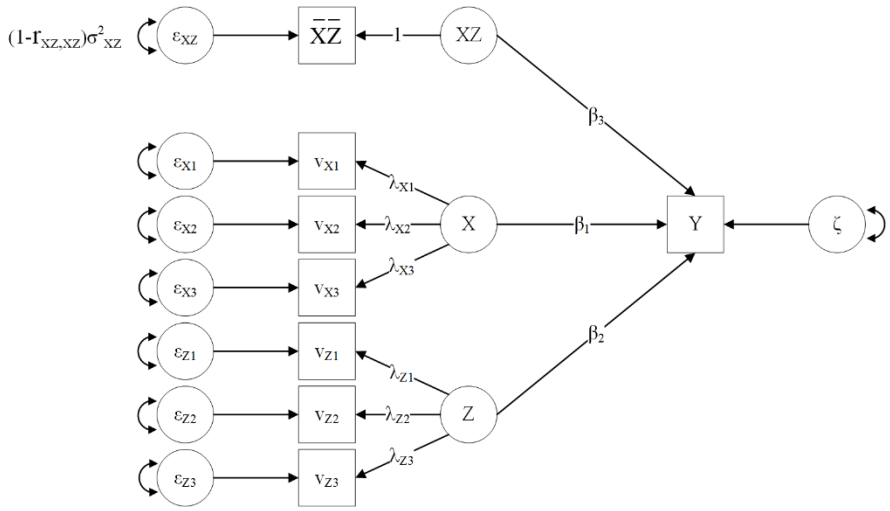
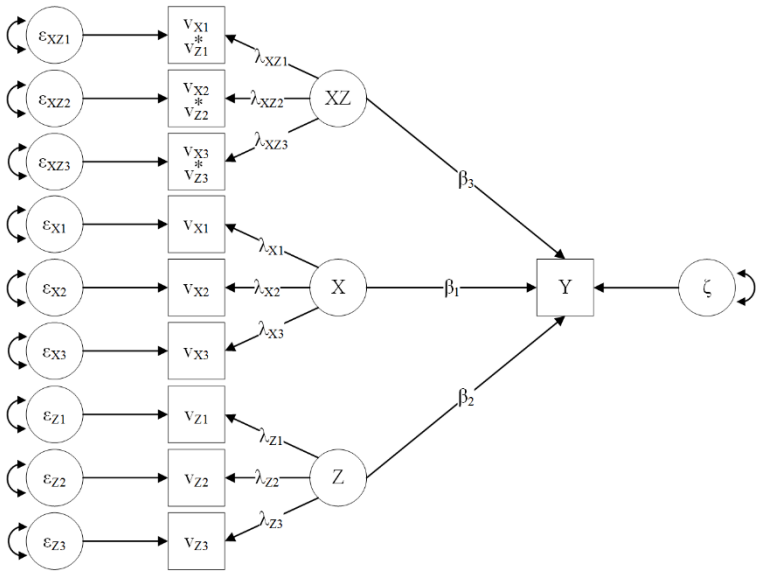
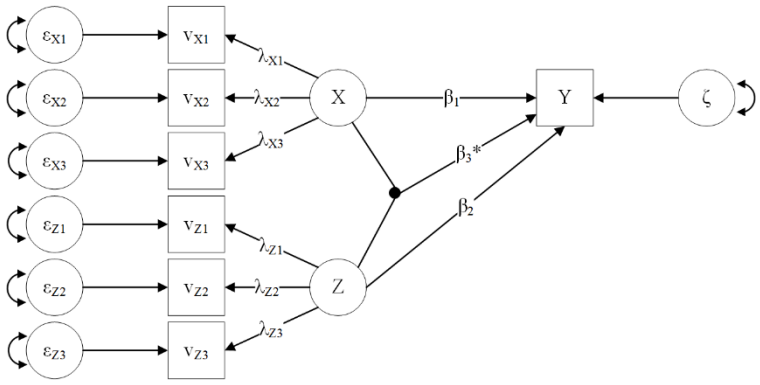


Figure 3.1 (CONTINUED)

Panel F: 2.3 Product indicators



Panel G: 2.4 Latent product



Notes: Circles are latent variables, and boxes are observed variables. Bars (e.g., $\overline{X}$) denote means, and hats (e.g., $\widehat{X}$) denote estimated factor scores. Unidirectional arrows refer to loadings and regression paths, and bidirectional arrows refer to error variances. Covariances between explanatory variables X, Z, and XZ are omitted for brevity. βs are regression coefficients, λs are factor loadings, ζs are structural error terms, and εs are measurement errors. Superscript g refers to a categorical grouping variable and subscript j to the indicator. $\tilde{Z}$ is the median of Z. σ^2_{XZ} is the variance of the interaction term XZ, and $r_{XZ,XZ}$ refers to its reliability, which can be estimated with Equation 3.4. The dot connecting X and Z refers to the notion that the moderation effect β^*_3 is inferred from the joint indicator distribution and not from an observed product of X and Z.

Panel A in Figure 3.1 depicts the framework which derives and compares the six moderation analysis methods. Table 3.1 summarizes and compares all six methods. Appendix 3A illustrates hypothetical data. We now discuss each method in turn and compare them.

3.2.2 Six methods for moderation analysis.

Method 1.1: Means.

This method uses the raw unweighted means (or sum), denoted $\overline{X}$ and $\overline{Z}$, of the three indicators of X and Z. To facilitate interpretation and reduce unessential multicollinearity, $\overline{X}$ and $\overline{Z}$ are mean-centered prior to computing the interaction term $\overline{X}\overline{Z}$ (Cohen et al. 2003). Panel B in Figure 3.1 illustrates the method. A test of β_3 against zero is a test of moderation.

The use of Method 1.1 (Means) assumes that indicators are measured without error or that the error is ignorable. Violating this assumption biases the estimates downward (Blalock 1965; Greene 2008; Spearman 1904). The estimated moderation effect $\hat{\beta}_3$ when not accounting for measurement error is (Greene 2008):

$$\hat{\beta}_3 = \beta_3 \times r_{XZ,XZ}, \tag{3.3}$$

where β_3 is the true moderation effect, and $r_{XZ,XZ}$ is the reliability of the interaction variable XZ. When XZ contains measurement error, its reliability is less than one, and the estimated moderation effect is biased downward toward 0. For example, when XZ has a reliability of .80, $\hat{\beta}_3$ is biased downward by 20%.

Method 1.2: Multi-group.

Method 1.2 splits the sample into G subgroups that differ in their value on the moderator $\overline{Z}$ and estimates a multi-group model (Baron and Kenny 1986; Goldsby et al. 2013). The multi-group model does not contain an interaction between X and Z but estimates a β_1 parameter for each group. The main effect of Z is in the intercept of Y^g, and a formal test for moderation assesses difference between models with and without moderation ($\beta_1^g = \beta_1$). Panel C in Figure 3.1 offers a visual representation. Grouping is straightforward when Z consists of

generally recognized subgroups, such as different countries, consumers of different brands, and so on. Yet grouping also requires discretization based on a median or other split when Z is continuous, as is the focus in this research.

Method 2.1: Factor scores.

Method 2.1 is a regression of factor scores $\hat{X}$ and $\hat{Z}$, extracted from a factor analysis, and $\hat{X}\hat{Z}$ is the product of these factor scores. Factor scores are estimates of the unobserved latent variables for the individual observations, with two benefits over means (Method 1.1). First, factor analysis decomposes the variance of an indicator into systematic variance that contributes to the latent variable (λ in Equation 3.2) and measurement error (ε in Equation 3.2). Second, it allows a different weight for each indicator of the latent variables, in that it includes estimates of the factor loadings (λ in Equation 3.2). Panel D in Figure 3.1 gives a visual representation, and a test of β_3 against zero provides the test of moderation.

Method 2.2: Corrected means.

Method 2.2 specifies a product of means $\overline{X}\overline{Z}$, as in Method 1.1 (Means), but accounts for its measurement error by using an estimate of the measurement reliability of the interaction variable XZ. Statistically, the interaction variable here is a single indicator of a latent variable with its loading fixed to $\lambda_{XZ} = 1$ and its error variance fixed for identification to $\sigma^2_{\varepsilon_{XZ}} = \left(1 - r_{XZ,XZ}\right)\sigma^2_{XZ}$, where σ^2_{XZ} is the variance of $\overline{X}\overline{Z}$, and $r_{XZ,XZ}$ is its reliability (Bollen 1989; Cole and Preacher 2014; Jöreskog and Sörbom 1989). Thus, this method accounts for measurement error, but when multiple indicators for X and/or Z are combined into a single composite score, it assumes equal weights of the indicators (tau equivalence). The estimate of the reliability of the interaction term (XZ) is:

$$r_{XZ,XZ} = \frac{r_{X,X} \times r_{Z,Z} + r^2_{X,Z}}{1 + r^2_{X,Z}}, \tag{3.4}$$

where $r_{X,X}$ is the reliability of X, $r_{Z,Z}$ is the reliability of Z, and $r^2_{X,Z}$ is the squared observed correlation between X and Z (Busemeyer and Jones 1983). Panel E in Figure 3.1 summarizes the method (also see Ping 1995). A test of β_3 against zero is a test of moderation.

Method 2.3: Product indicators.

Method 2.3 uses product indicators that load on a latent interaction variable XZ (Marsh et al. 2004), as summarized in Panel F of Figure 3.1. There are various approaches to constructing product indicators.[3] A common one applies double mean-centering to the indicators (Lin et al. 2010) and uses "matched pairs" (Marsh et al. 2004) of all indicators of X and Z, but only once. Thus in our example, it would result in three product indicators: $v_{X1}\times v_{Z1}$, $v_{X2}\times v_{Z2}$, and $v_{X3}\times v_{Z3}$. A test of β_3 against zero is the test of moderation.

Method 2.4: Latent product.

The latent product method estimates the moderation effect from the latent product of X and Z (Klein and Moosbrugger 2000). Panel G in Figure 3.1 shows the method. The dot in the figure and the estimation of β_3^* instead of β_3 reflect that the latent product method estimates the moderation effect from the product of latent X and latent Z. A test of β_3^* against zero is a

[3] The product indicator approach was introduced by Kenny and Judd (1984) after which others continued its development (Algina and Moulder 2001; Jaccard and Wan 1995; Jöreskog and Yang 1996; Kenny and Judd 1984; Lin et al. 2010; Little et al. 2006; Marsh et al. 2004). Marsh et al. (2004) and Cortina et al. (2019) have overviews and R code. This chapter focuses on the version that double mean-centers matched pairs of indicators (Lin et al. 2010; Marsh et al. 2004) for several reasons. First, the matched pairs double mean-centering approach is accessible in that it does not require constraints on the measurement model. Other approaches require non-linear constraints that are cumbersome, error prone, and not readily available in statistical software packages. Second, the use of the other product indicator approaches has been limited. To illustrate this, we performed a citation search using Web of Science that found 4 articles that contain citations of Kenny and Judd (1984), 2 of Jaccard and Wan (1995) one of Jöreskog and Yang (1996), none of Algina and Moulder (2001) and Little et al. (2006), and 3 of Marsh et al. (2004), published in *Journal of Marketing* and *Journal of Marketing Research*. Yet only one citing article (Lusch and Brown 1996) contained an application that used the Kenny and Judd (1984) indicators and constraints. Cortina et al. (2019, p. 6) reached a similar conclusion and noted that "...none of the 562 authors who cite Jöreskog and Yang (1996; as of November 2018, Google Scholar) do so because they actually use the procedure." The remainder of the citing articles that contained applications, as well as all moderation tests reported in Table 3.2, used the matched pairs double mean-centering approach. Third, earlier simulation studies showed that the unconstrained method performed equally well as various constrained methods (Marsh et al. 2004). Hence, this chapter focuses on the more accessible and more common matched pairs double mean-centering method (Lin et al. 2010; Marsh et al. 2004).

test of moderation. The method uses the full information in the raw data and does not use product indicators like Method 2.3 does.

Instead of product indicators to specify the latent interaction, the latent product method relies on the non-normal indicator distribution f(v, Y), which is represented by a weighted sum or mixture of normal distributions (Klein and Moosbrugger 2000). Formally:

$$f(v, Y) = \sum_{j=1}^{M} \rho_j \varphi_{\mu_j, \Sigma_j}(v, Y), \tag{3.5}$$

where j ϵ (1, 2, …, M) denotes the mixture components, ρ_j are the weights, and φ_{μ_j,Σ_j} is the multivariate normal distribution. Model estimation uses an expectation maximization algorithm. Then, the mean and covariance matrices (μ_j and σ_j) implied by the model in Equations 3.1 and 3.2 get entered into Equation 3.5. In practice, M is fixed to 16, which generally is sufficient to describe a single continuous moderation effect. Klein and Moosbrugger (2000) provide technical details.

3.2.3 Comparison of the six methods.

The comparison of the six moderation analysis methods in Table 3.1 indicates whether the methods account for measurement error to estimate the moderation effect. We also outline some strengths and weaknesses and provide illustrative applications in marketing research. Methods 1.1 (Means) and 1.2 (Multi-group) are straightforward to apply. It is true that a multi-group model might account for measurement error in X, and that a mean score of a multi-indicator scale typically has a higher reliability than using single-indicators.[4] Yet, the methods do not account for the unreliability of the composite to estimate the moderation effect. The multi-group method, while appropriate for naturally categorical moderators, also

[4] This can be shown with the formula of standardized Cronbach's alpha: $\frac{k\bar{r}}{(1+(k-1)\bar{r}}$, where k is the number of indicators of a multi-indicator measure, and $\bar{r}$ is the average correlation between the indicators, assuming all are equally good. For instance, if three indicators of a construct that are correlated .50 have a single-indicator reliability of .50, the multi-item reliability is .75. Chapter 5 returns to single-indicator reliability.

requires discretizing continuous moderators, which adds even more unreliability and uses only partial information in Z, leading to bias and lower power (Maxwell and Delaney 1993). Example applications of the means and multi-group methods are available in Mende et al. (2013) and Homburg et al. (2008), respectively.

The four other methods more adequately account for measurement error. They account for unreliability in the indicators by decomposing the variance in systematic variance and error variance. Although they should be unbiased in sufficiently large samples, if measurement error is adequately accounted for, it is not apparent how they compare in terms of bias and power in various conditions, which we attempt to address subsequently with our Monte Carlo simulations. Method 2.4 (Latent product) uses all the information in the raw data and simultaneously estimates the factor loadings with the moderation effect. This method thus should have the lowest bias and highest power. Yet it is computationally intensive, because the estimation algorithm requires numerical integration. Nor is it available in standard statistical software packages, with the exceptions of its implementations in Mplus (Muthén and Muthén 2018) and a dedicated package in R (Umbach et al. 2017). With this method, Korschun et al. (2014) find that the extent to which frontline employees identify with the organization and customers depends on how much the employees perceive that managers and customers support the company's corporate social responsibility activities. The effects were stronger for employees to whom corporate social responsibility activities are important.

Method 2.1 (Factor scores) computes the product of latent variables X and Z to estimate the moderation effect, similar to Method 2.4 (Latent product). Voss and Voss (2000) apply this method in their study of the moderating effects of interfunctional coordination on the impacts of product, competitor, and customer orientations on firm performance. Using factor scores can be viewed as a two-step estimation of moderation effects by estimating the measurement and structural parts of the model separately (Anderson and Gerbing 1988). The

method first estimates the factor scores for X and Z and calculates their product, then estimates the moderation effect of XZ. While factor scores have theoretical indeterminacy, they are estimates of the latent variables (Grice 2001; Lastovicka and Thamodaran 1991; McDonald and Burr 1967; Tucker 1971). Importantly, and unlike the scores used for the means method, they are estimated from a model that decomposes the variance in X and Z in systematic variance and error variance, and freely estimates the weights of each indicator to the latent variable. We use a common regression-based factor score, which has been shown to work well in settings similar to ours (Devlieger et al. 2016; Lastovicka and Thamodaran 1991; Lu et al. 2011; Ng and Chan 2020; Skrondal and Laake 2001). Although the method should thus be able to recover the unstandardized moderation effect (Skrondal and Laake 2001), because properly estimated factor scores have the same disattenuating properties as latent variables, it might have less statistical power than Method 2.4 (Latent product) due to its two-step estimation. A disadvantage of the method could be that the factor score itself is more difficult to interpret than a mean due to its indeterminacy and scale (Grice 2001; Lastovicka and Thamodaran 1991).

Method 2.2 (Corrected means) accounts for measurement error in either single-indicator measures, or composites of multi-indicator measures as single-indicators. It requires an estimate of the single-indicator reliability. It uses partial information when multi-indicator scales are available, and relies on their unweighted means to estimate the moderation effect. Accordingly, it should have less power than Method 2.4 (Latent product), which uses all information in the raw data. De Luca and Atuahene-Gima (2007) apply Method 2.2 in research on the effects of market knowledge dimensions and cross-functional collaboration on firms' product innovation performance, as moderated by knowledge integration mechanisms. Finally, Method 2.3 (Product indicators) simultaneously estimates the measurement model with the moderation effect. However, the matched pairs approach to

Table 3.1

Comparison of Six Methods for Moderation Analysis

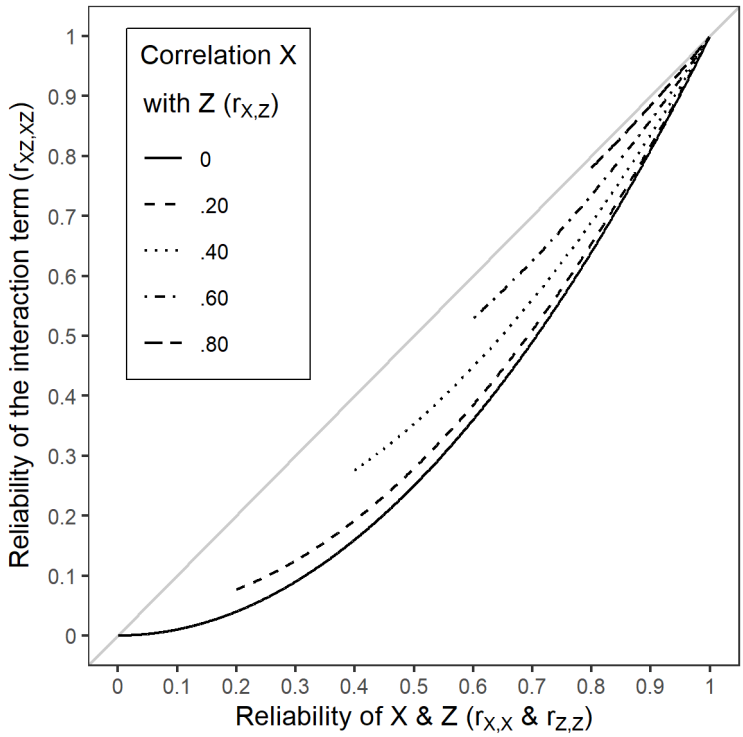
Method	Description	Account for measurement error to estimate the moderation effect	Other strengths (+) and weaknesses (-)	Illustrative application
1.1 Means	Product of unweighted means of the indicators	No	+ Not complex in application and widely available	Mende et al. (2013)
1.2 Multi-group	Estimate separate model per group, based on discretization of the continuous moderator	No	+ Not complex in application and widely available - Discretization of a continuous moderator Z uses partial information in Z and adds additional measurement error to Z	Homburg et al. (2008)
2.1 Factor scores	Product of factor scores of the latent variables	Yes	+ Not complex in application and widely available - Does not simultaneously estimate the factor analysis and moderation effect - Factor scores might be difficult to interpret	Voss and Voss (2000)
2.2 Corrected means	Product of unweighted means of the indicators, corrected for measurement error	Yes	+ Can be used to account for unreliability in single-indicator measures - Requires a fixed reliability estimate of the interaction term	De Luca and Atuahene-Gima (2007)
2.3 Product indicators	Products of pairs of indicators	Yes	+ Simultaneously estimates the factor analysis and moderation effect - Requires selecting product indicator pairs - Matched pairs approach uses partial information in the indicators	Homburg et al. (2013)
2.4 Latent product	Product of latent variables	Yes	+ Uses full information in the raw data + Simultaneously estimates the factor analysis and moderation effect - Limited availability and computationally intensive	Korschun et al. (2014)

select product indicators includes partial information, because it uses each indicator only once (Marsh et al. 2004), so it may have less power than Method 2.4 (Latent product). The low reliability of multiplications of indicators with low reliabilities themselves might also lead to model convergence issues. However, it is not apparent a priori how it compares with Methods 2.1 (Factor scores) and 2.2 (Corrected means). Homburg et al. (2013) adopt this method to estimate the effects of various moderators (e.g., market-related uncertainty, competition intensity) on the relationship among business practice, corporate social responsibility reputation, and trust.

3.3 The Effect of Multicollinearity on the Bias and Power of the Moderation Effect

The regression equation needs to include the main effects of X and Z to test the moderation effect of XZ appropriately (Cohen et al. 2003). It is common in non-experimental research for

Figure 3.2
The Reliability of the Interaction Term Increases When the Correlation Between X and Z Increases



Notes: This figure plots the reliability of the interaction term as a function of the reliability of X and Z, for different observed correlations between X and Z ($r_{X,Z}$), using Equation 3.4. The greyed 45° line indicates the situation in which the reliability of the interaction term would increase proportionally with the reliability of X and Z.

X, Z, and XZ to be correlated, which leads to multicollinearity in the regression equation. Measurement error in X and Z can mask high correlations between them, and accounting for measurement error can make these high correlations emerge (Grewal et al. 2004). Multicollinearity increases the standard errors of the parameter estimates (Greene 2008), even when measurement error is accounted for (Grewal et al. 2004).

Ironically, multicollinearity between X and Z reduces the bias of the moderation effect when measurement error is unaccounted for because multicollinearity increases the reliability of the interaction term XZ. To depict this point, Figure 3.2 plots the reliability of the interaction term XZ (Equation 3.4) for reliabilities of X and Z ($r_{X,Z}$) between 0 and 1 and for observed correlations of X and Z from 0 to 1 in steps of .20. For example, when the reliability of X and Z is .80, the reliability of the interaction term (XZ) is .64 for $r_{X,Z} = 0$, and it is .69 for $r_{X,Z} = .40$. The maximum reliability of the interaction term occurs when X and Z are maximally correlated. The highest observable correlation between two variables is, at most, the square root of the product of the reliabilities: $r_{X,Z}^{observed} = r_{X,Z}^{true} \times \sqrt{r_{X,X} \times r_{Z,Z}}$ (Spearman 1904). The lines in the figure in turn are truncated; high correlations can only be observed when measures are reliable. Intuitively, when the reliability of X and Z increases, the reliability of XZ increases, but its reliability is always below the 45-degree line, except at the extremes. In other words, the reliability of XZ is lower than the reliability of X and Z when X and Z have the same reliability.

Thus, on the one hand, multicollinearity increases the reliability of the interaction term, such that less measurement error needs to be accounted for, which increases power. On the other hand, multicollinearity increases standard errors, which lowers power. In turn, we need to compare the performance of the six moderation analysis methods at varying levels of multicollinearity between X and Z. We also return to the net effect of multicollinearity on the statistical power of the moderation effect in our Monte Carlo simulation.

3.4 Literature Review of the Six Moderation Analysis Methods

The purpose of the literature review is twofold. First, we seek to assess the practice of using the six moderation analysis methods in marketing research. Second, the results serve as an input for the Monte Carlo simulations we use to assess the performance of the six methods. We searched all articles published in *Journal of Marketing (JM)* and *Journal of Marketing Research (JMR)* between 2000 and 2017 for keywords related to latent variables and moderation. Specifically, the search was ((factor analysis) OR (measurement model) OR (factor score)) AND (moderation OR moderate OR moderator OR moderates OR moderated OR moderating OR contingency OR interaction OR interact). We identified 276 *JM* and 167 *JMR* articles. We checked the citations of articles that proposed specific moderation methods and did not find additional eligible articles. We selected studies in which both X and Z are measured with multiple observed indicators; we do not consider moderation with categorical or manipulated variables, as is common in experimental research. For an effect to be included in the analysis, it must contain at least one hypothesized two-way interaction between two latent variables, measured with at least three (semi-)continuous indicators (e.g., on 7-point scales). The dependent variable could have any number of indicators. We excluded effects estimated with partial least squares, three-way interactions, and (moderated) polynomials, to ensure that the moderation effects of interest are comparable. Ultimately, we identified 97 (78 in *JM* and 19 in *JMR*) articles published between 2000 and 2017 that theorized and tested 504 (427 in *JM* and 77 in *JMR*) moderation effects, for an average of 5.20 (median = 4, SD = 6.13, range = 1–48) effects per article.

How widely are the six moderation analysis methods used? Table 3.2 shows that 89%, or the vast majority of the 497 moderation tests for which the method could be unequivocally determined, used Method 1.1 (Means; n = 387, 78%) or Method 1.2 (Multi-group; n = 56, 11%). The majority (n = 39, 93%) of 42 multi-group tests for which the split could be

Table 3.2
Moderation Analysis Methods in *Journal of Marketing* and *Journal of Marketing Research* 2000-2017

Method	Frequency (% out of 497)	Frequency of significant moderation effects
1. Moderation methods that do not account for measurement error		
1.1 Means	387 (78%)	211 (55% of 387)
1.2 Multi-group	56 (11%)	41 (73% of 56)
	443 (89%)	252 (57% of 443)
2. Moderation methods that do account for measurement error		
2.1 Factor scores	18 (4%)	4 (22% of 18)
2.2 Corrected means	18 (4%)	11 (61% of 18)
2.3 Product indicators	13 (3%)	10 (77% of 13)
2.4 Latent product	5 (1%)	5 (100% of 5)
	54 (11%)	30 (56% of 54)
Total amount of moderation effects	497	282 (57% of 497)

Notes: The method could be unequivocally determined from the study description for 497 out of 504 moderation effects. Percentages may not add to 100% due to rounding.

Table 3.3
Properties of the Data of Moderation Analyses in *Journal of Marketing* and *Journal of Marketing Research* 2000-2017

Category	Result
Sample size per study (504 out of 504 effects)	M = 587, Mdn = 171 (SD = 1,838, range = 47-14,994)
Number of indicators for Y (504 out of 504 effects)	M = 4.21, Mdn = 3 (SD = 3.55, range = 1-17)
Reliability of Y (307 out of 504 effects)	M_w = .86, M = .87, Mdn = .85 (SD = .06, range = .61-.96)
Number of indicators for X and Z (1,008 out of 1008 variables)	M = 5.47, Mdn = 4 (SD = 4.39, range = 3-40)
Reliability of X and Z (910 out of 1008 variables)	M_w = .86, M = .85, Mdn = .86 (SD = .08, range = .48-.99)
Absolute correlation of X with Z (422 out of 504 effects)	M_w = .25, M = .26, Mdn = .20 (SD = .19, range = 0-.78)
Absolute correlation of X and Z with Y (812 out of 1008 variables)	M_w = .24, M = .23, Mdn = .19 (SD = .16, range = 0-.77)
Absolute correlation of XZ with Y (22 out of 504 effects)	M_w = .17, M = .17, Mdn = .19 (SD = .10, range = .01-.33)

Notes: The numbers in parentheses in the first column denote the number of effects or variables that the corresponding statistics in the remaining columns are based on because some statistics could not always be unequivocally determined from study descriptions. M_w is the weighted (by $\sqrt{n-2}$) mean, M is the (simple) mean, Mdn is the median, and SD is the standard deviation. Means and weighted means of reliabilities and correlations are based on back-transformed Fisher-Z transformed values (Charter and Larsen 1983).

unequivocally determined discretized Z with a median or mean split. Eighteen (4%) effects used Method 2.1 (Factor scores), and 18 (4%) used Method 2.2 (Corrected means). Thirteen cases (3%) used Method 2.3 (Product indicators), and five (1%) used Method 2.4 (Latent product).

Thus, most moderation tests in our review do not account for measurement error. We also coded whether the authors found that the tested moderation effects were statistically significant. Approximately 57% of the 443 moderation tests that accounted for measurement error, and 56% of the 54 tests that did not, were reported to be statistically significant. These proportions do not statistically differ (z-statistic = .19, p = .85).

We also assessed the properties of the data used to estimate the moderation effects, as well as the size of the moderation effects (see Table 3.3). We used meta-analysis to determine the mean reported reliabilities and correlations. We transformed these to Fisher-Z-values, took the mean, and back-transformed it to a meta-analytic mean correlation or reliability (Charter and Larsen 1983). We report a simple and weighted mean, which uses the inverse of the standard error of the Z-values $\sqrt{\mathrm{n}-2}$ as a weight, giving more weight to correlations from larger studies.

The median sample size is 171, remarkably close to the median of 178 determined in an early review of structural equation models in marketing (Baumgartner and Homburg 1996) and the mean of 183 in a more recent review of mediation analyses in consumer research (Pieters 2017). The median number of indicators for Y is 3, but Y is commonly measured with a single indicator (n = 147, 29% of 504). The median number of indicators for X and Z is 4. The median reliabilities are .85 for Y and .86 for X and Z, in line with recent findings (Pieters 2017) but higher than the mean reliability of .77 revealed in an early meta-analysis of measurement reliability (Peterson 1994).

We extracted correlations between the measures from the correlation tables in the articles. The weighted mean absolute correlation between X and Z (multicollinearity) is .25. The weighted mean correlation between the interacting variables (X and Z) and Y is .24, in line with the mean effect size of .24 in a meta-analysis of meta-analyses in marketing (Eisend 2015). The weighted mean correlation of XZ with Y could only be assessed for 22 moderation effects, because most articles do not report correlations of interaction terms; it is .17. Although the mean correlation of X and Z with Y is not statistically significant from the correlation of XZ with Y for this modest number of effects, the magnitudes are consistent with the conventional wisdom that moderation effect sizes are smaller than main effects (Aguinis et al. 2005; Eisend 2015).

We next use the findings from the literature review in Monte Carlo analyses to determine the impact of the six moderation analysis methods on the bias and power of the moderation effect.

3.5 Performance of the Six Methods

3.5.1 Method.

We conducted Monte Carlo simulations to assess the performance of the six moderation methods under controlled conditions. We systematically varied the properties of the data in terms of the sample size (n), reliability ($r_{X,X}$ & $r_{Z,Z}$), effect size of the moderation effect (β_3), effect size of the main effects (β_1 & β_2), and correlation between X and Z ($r_{X,Z}$). We based the design on the results of the literature review. The sample sizes were fixed to 50, 100, 150, 175, 200, 300, 400, 500, 750, and 1,500. About 94% of the estimated effects in the literature review had samples between 50 and 1,500 observations. We used smaller increments between 50 and 200 observations, because 54% of the effects in the literature review were tested with samples in that range. We fixed the reliabilities of X and Z to .70 (commonly considered a lower bound of acceptable reliability), .80 (good), and .90 (excellent) (Peterson 1994). The

percentage of X and Z variables in the literature review that achieved reliability between .70 and .90 was about 78%. We used the median effect size in the literature review, .20, plus and minus approximately one standard deviation, so about .15 for the main effects and .10 for the moderation effect. Thus, the true main effects (β_1 and β_2) were .05, .20, and .35. Only 12% of the main effects reported had a main effect smaller than .05, and only 20% were larger than .35. The true moderation effects (β_3) were .10, .20, and .30. Only one investigated moderation effect had an effect size larger than .30, and 32% had an effect size smaller than .10. Although the moderation effect sizes are at the high end of the range of observed effect sizes, they are reasonable, because reported correlations are commonly attenuated, and interaction terms have relatively low reliabilities (cf. Equation 3.4). We fixed the correlation between X and Z to 0, .20, .40, and .60, because in the literature review, 95% of the effects showed correlations covered by those values. In summary, our full-factorial design contains 10 (sample size) × 3 (reliability) × 3 (effect size moderation effect) × 3 (effect size main effects) × 4 (correlation X with Z) = 1,080 cells. We generated 5,000 replications for each cell in the design, resulting in 5,400,000 data sets.

Data generation and analysis were conducted in R (R Core Team 2019). Following Devlieger et al. (2016), we generated the data in two steps. First, we obtained X and Z from a multivariate standard normal distribution, varying their correlation and the sample size according to the design, and XZ is the product of the latent X and Z variables. For Y, we used Equation 3.1, and the residual variance was fixed to 1. Second, Equation 3.2 generated three indicators for X and Z. We followed Grewal et al. (2004) and fixed the loadings to one and the measurement error variances according to the reliability in the design. With standardized X and Z, each with three indicators, the measurement error variances were: $3 \times (1 - r_{X,X})/r_{X,X}$ (Grewal et al. 2004). Thus for a reliability of X and Z of .70, error variances were fixed to 1.29; the error variances were .75 and .33 for reliabilities of .80 and .90, respectively.

We apply the six moderation methods to analyze the data. Method 1.1 (Means) mean-centers X and Z prior to creating the interaction term (Cohen et al. 2003). Method 1.2 (Multi-group) uses a median split as a grouping variable. Method 2.1 (Factor scores) estimates a two-factor (X and Z) confirmatory factor analysis in the first step, then extracts regression factor score estimates (Devlieger et al. 2016). Method 2.2 (Corrected means) estimates a two-factor confirmatory factor analysis of X and Z to obtain estimates of composite reliabilities (Fornell and Larcker 1981), used to determine the reliability of the interaction term XZ with Equation 3.4. Method 2.3 (Product indicators) follows a matched pairs recommendation by Marsh et al. (2004) and double mean-centers the product indicators (Lin et al. 2010). We use the nlsem package (Umbach et al. 2017) in R for Method 2.4 (Latent product) and the lavaan package (Rosseel 2012) for the other methods. All methods rely on standard maximum likelihood estimation except Method 2.4 (Latent product), which uses the expectation maximization algorithm. The data generation for the full-factorial design and estimation of the six methods required a cluster of 72 Intel Xeon processors at 2.60 GHz, running for about four consecutive days. Appendix 3B details the estimated bias and power of the moderation effects.

3.5.2 Results.

Bias of the moderation effect.

Column A in Table 3.4 summarizes the results of a meta–analysis of variance (ANOVA) pertaining to the bias of the moderation effects. The factors and their two-way interactions jointly account for almost 98% of the variance. We find substantial differences in the bias across methods, accounting for about 82% of the variance; these differences also depend on the estimation sample size (about 2% variance accounted for by the interaction). The interaction of the method and the reliability of X and Z also accounts for about 10% of the variance in bias, suggesting the effect of reliability on bias differs across methods.

To understand these differences in more detail, we also plot the bias of the moderation effects as a function of the sample size and the reliability of X and Z. The left side of Figure 3.3 reveals the bias of the moderation effects for the six methods over the logged sample size. Neither Methods 1.1 (Means) nor 1.2 (Multi-group) provide consistent estimates of the moderation effects and have large downward biases of about -32% and -36%, respectively. The four methods that account for measurement error exhibit much smaller biases and consistent estimates with sufficiently large sample sizes. Bias is lowest for Method 2.1 (Factor scores) and Method 2.4 (Latent product), at about 1%, for a sample size of 175 (approximately the median, according to the literature review). Method 2.2 (Corrected

Table 3.4
Bias and Power of the Moderation Effect: Variance Accounted for

		(A) Bias		(B) Power	
Factor	d.f.	% var. accounted for	F (*p*-value)	% var. accounted for	F (*p*-value)
Main effects					
Method (M)	5	**81.79**	46945 (< .001)	**3.38**	2049 (< .001)
Sample size (n)	9	**1.28**	408 (< .001)	**43.30**	14596 (< .001)
Reliability X and Z ($r_{X,X}$ & $r_{Z,Z}$)	2	.78	1124 (< .001)	**2.18**	3310 (< .001)
Effect size moderation effect (β_3)	2	.30	436 (< .001)	**41.45**	62877 (< .001)
Effect size main effects (β_1 & β_2)	2	.05	68 (< .001)	.05	70 (< .001)
Correlation between X and Z ($r_{X,Z}$)	3	.07	65 (< .001)	.85	861 (< .001)
Two-way interaction effects					
M × n	45	**2.16**	137 (< .001)	.34	23 (< .001)
M × $r_{X,X}$ & $r_{Z,Z}$	10	**9.97**	2862 (< .001)	.50	152 (< .001)
M × β_3	10	.03	7 (< .001)	.07	22 (< .001)
M × β_1 & β_2	10	.23	66 (< .001)	.04	11 (< .001)
M × $r_{X,Z}$	15	.48	92 (< .001)	.03	6 (< .001)
n × $r_{X,X}$ & $r_{Z,Z}$	18	.45	72 (< .001)	.16	27 (< .001)
n × β_3	18	.09	15 (< .001)	**5.43**	915 (< .001)
n × β_1 & β_2	18	.01	2 (= .018)	.00	0 (= .999)
n × $r_{X,Z}$	27	.02	2 (< .001)	.06	7 (< .001)
$r_{X,X}$ & $r_{Z,Z}$ × β_3	4	.07	51 (< .001)	.04	27 (< .001)
$r_{X,X}$ & $r_{Z,Z}$ × β_1 & β_2	4	.01	5 (= .001)	.01	7 (< .001)
$r_{X,X}$ & $r_{Z,Z}$ × $r_{X,Z}$	6	.02	8 (< .001)	.01	6 (< .001)
β_3 × β_1 & β_2	4	.00	1 (= .646)	.00	1 (= .390)
β_3 × $r_{X,Z}$	6	.00	1 (= .219)	.03	15 (< .001)
β_1 & β_2 × $r_{X,Z}$	6	.00	2 (= .058)	.00	0 (= .887)
Residual	6255	2.18		2.06	

Notes: n = 6,480 (1,080 cells × 6 methods). The results are based on a meta-ANOVA with the design factors of the Monte Carlo simulation as main effects, and all two-way interactions. The dependent variable in column (A) is the bias of the moderation effect, and the dependent variable in column (B) is the power of the moderation effect (see Appendix 3B for details). Table entries are degrees of freedom (d.f.), % variance accounted for, and F-statistics with *p*-values. Percentages may not add to 100% due to rounding; percentages higher than 1% are in bold.

means) has a small upward bias of about 4%, and Method 2.3 (Product indicators) invokes a bias of 12%, which drops to 1% when the sample size is very large, such as 1,500.

The right-hand plot in Figure 3.3 depicts the bias for different levels of reliability of X and Z. Increasing the reliability of X and Z generally decreases the bias, except in the cases of Methods 2.1 (Factor scores) and 2.4 (Latent product), which adequately account for measurement error, regardless of its magnitude. At a reliability of .70, Methods 2.2 (Corrected means) and 2.3 (Product indicators) are moderately biased, by about 7% and 20%, which falls to about 2% when the reliabilities are .90. Even when X and Z achieve reliability of .90, a common threshold for excellent reliability, Methods 1.1 (Means) and 1.2 (Multi-group) remain severely biased, by about 17% and 27%. Thus, the reliability of the interaction term XZ is relatively low, compared with the reliabilities of the interacting variables X and Z, especially when they are uncorrelated (Equation 3.4).

Power of the moderation effect.

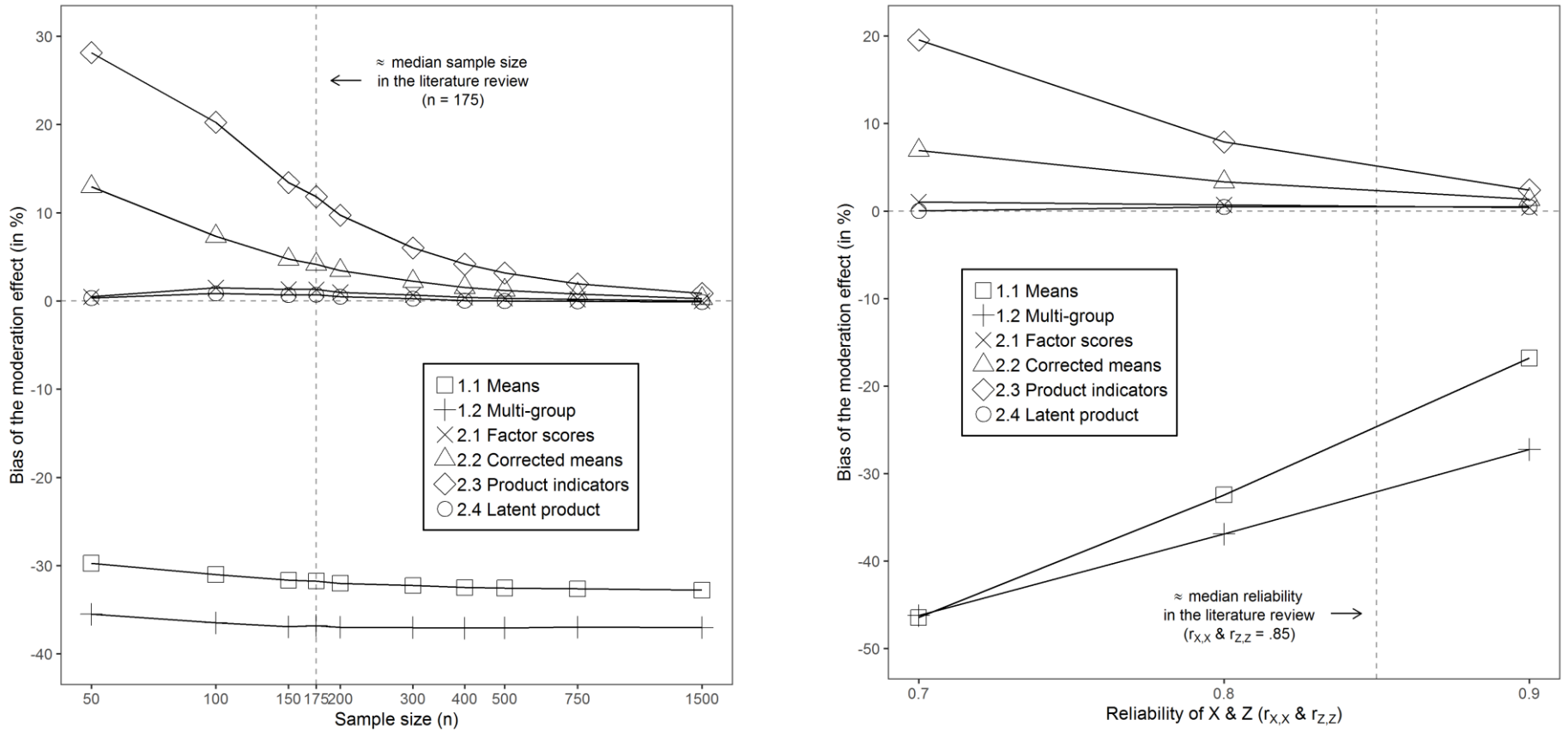
Column B in Table 3.4 reports the results of a meta-ANOVA on the statistical power of the moderation effect. The factors and their two-way interactions jointly account for almost 98% of the variance. The sample size (about 43%) and effect size of the moderation effect (about 41%) account for most of the variance in power of the moderation effect, with little difference across methods (respectively, .34% and .07% variance accounted for by the interactions). The correlation between X and Z has a modest effect (.85% variance accounted for) on the power of the moderation effect. The predicted means of the meta-ANOVA (full results not reported for brevity) further show that increasing the correlation between X and Z increases the power of the moderation effect. For example, increasing the correlation between X and Z from 0 to .60 increases the power for Method 1.1 (Means) from 63% to 71% and the power for Method 2.4 (Latent product) from 66% to 74%. Thus, multicollinearity has a

Figure 3.3

Impact of Sample Size and Reliability of X and Z on the Bias of the Moderation Effect

Panel A: Impact of sample size (n) on bias

Panel B: Impact of reliability of X and Z ($r_{X,X}$ & $r_{Z,Z}$) on bias



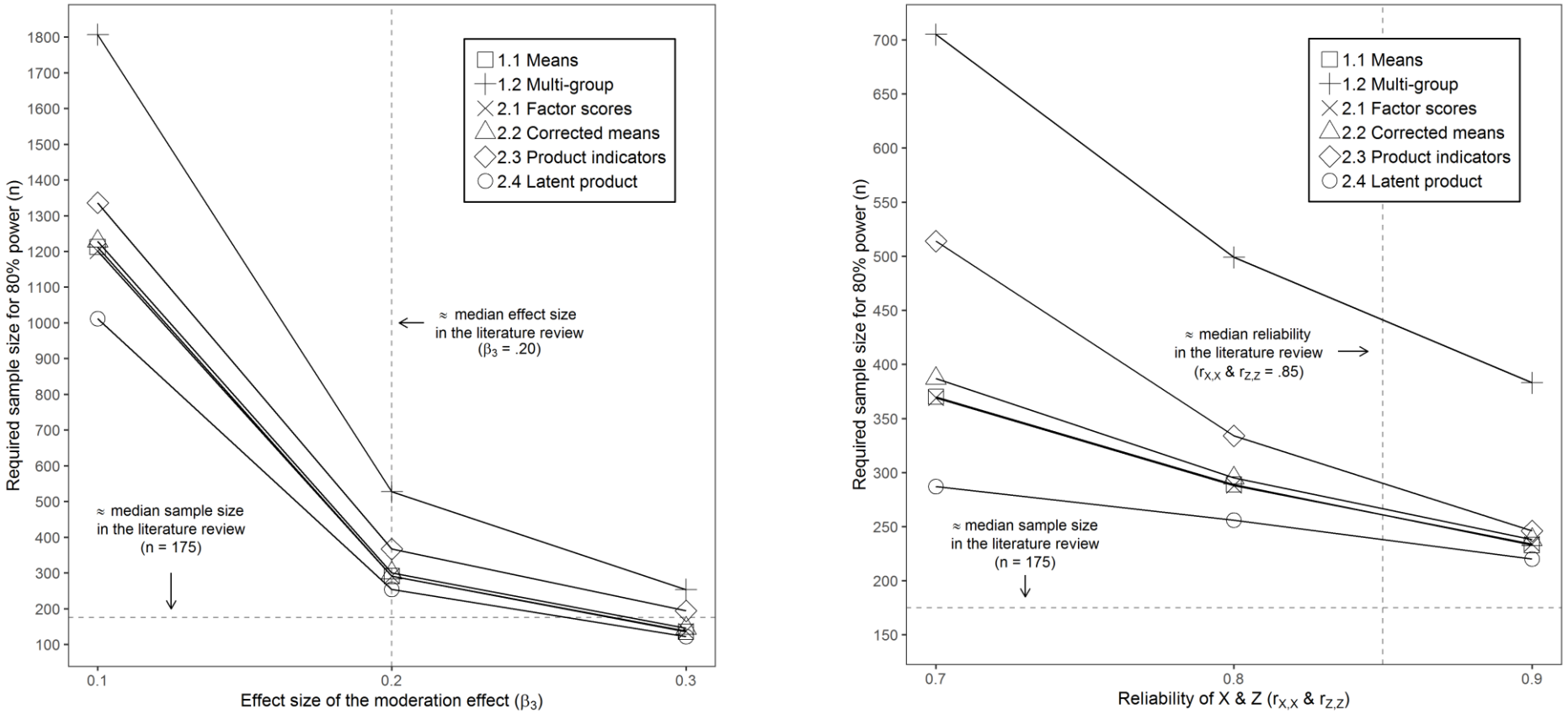
Notes: Icons denote predicted conditional means of bias of the moderation effect (in %) estimated by a meta-ANOVA with the Monte Carlo design factors and six moderation methods as main effects, and all two-way interactions (see Table 3.4). Solid lines are interpolations between the icons. Sample size is on a log-scale. Horizontal dotted lines in both plots indicate zero bias.

Figure 3.4

Required Sample Size for 80% Power as a Function of Effect Size of the Moderation Effect and Reliability of X & Z

Panel A: Required sample size (n) for 80% power

Panel B: Impact of reliability of X and Z ($r_{X,X}$ & $r_{Z,Z}$) on power



Notes: Left plot denotes the required sample size for 80% power to detect a significant moderation effect, across effect sizes of the moderation effect. Icons are interpolations of predicted conditional means of power of the moderation effect (in %), estimated by meta-ANOVAs with the Monte Carlo design factors and six moderation methods as main effects, and all two-way and three-way interactions. One sample size, n = 1,807, had to be extrapolated for Method 1.2 (Multi-group) because 80% power was not attained within the range of investigated sample sizes (50-1,500) for an effect size of .10. Right plot denotes the required sample size for 80% power to detect a significant true moderation effect of .20, across reliabilities of X and Z. Icons are predicted conditional means of power of the moderation effect (in %), estimated by a meta-ANOVA with the Monte Carlo design factors and six moderation methods as main effects, and all two-way, three-way, and four-way interactions. Solid lines in both plots are interpolations between the icons.

positive, if modest, effect on the power of the moderation effect. Furthermore, the interaction between the sample size and effect size of the moderation effect accounts for slightly more than 5% of the variance. The reliability of X and Z accounts for about 2% of the variance. Finally, the method accounts for 3% of the variance in the power of the moderation effect.

The left-hand plot in Figure 3.4 shows the required sample size for 80% power across the effect sizes of the moderation effect. As expected, estimating small moderation effects with 80% power requires larger samples than estimating large effects. The median moderation effect size in the literature review, about .20, needs bigger samples than 171, the median sample size, for 80% power. The best method in this respect is Method 2.4 (Latent product), which requires a sample size of 254. Methods 1.1 (Means), 2.1 (Factor scores), and 2.2 (Corrected means) need larger samples, of 291, 291, and 300, respectively. The largest required sample sizes are 367 for Method 2.3 (Product indicators) and 527 for Method 1.2 (Multi-group). Small effects of .10 require samples that are much larger: 1,212 (1.1 Means), 1,807 (1.2 Multi-group), 1,201 (2.1 Factor scores), 1,228 (2.2 Corrected means), 1,336 (2.3 Product indicators), and 1,012 (2.4 Latent product). Large effects of .30 have 80% power for 135 (1.1 Means), 253 (1.2 Multi-group), 137 (2.1 Factor scores), 145 (2.2 Corrected means), 194 (2.3 Product indicators), and 122 (2.4 Latent product) observations.

The right side of Figure 3.4 indicates the required sample size for 80% power of the moderation effect for different levels of reliability of X and Z, revealing a positive effect on the power of the moderation effect as they increase. For example, increasing the reliability from .70 to .90 decreases the required sample size for Method 1.1 (Means) from 370 to 233, the required sample size for Method 2.1 (Factor scores) from 369 to 234, and the required sample size for Method 2.4 (Latent product) from 287 to 220. The interpolated required sample sizes for 80% power at a reliability of .85, about the median reliability, are 261 (1.1

Means), 441 (1.2 Multi-group), 261 (2.1 Factor scores), 267 (2.2 Corrected means), 290 (2.3 Product indicators), and 238 (2.4 Latent product).

The modest effects of the analysis method (about 3%) and reliability of X and Z (about 2%) on the statistical power of the moderation effect arise because accounting for measurement error comes at a cost. That is, adequately accounting for measurement error recovers unbiased effects, but standard errors are smaller when there is less measurement error in the variables, which results into higher power levels (Grewal et al. 2004; Lomax 1986; Yuan et al. 2010). For illustration, we looked at the average standard error of the moderation effect when the sample size is 200, the sizes of the main and moderation effects are .20, and the correlation between X and Z is .20. Increasing the reliability of X and Z from .70 to .90 increases the average estimate for Method 1.1 (Means) from .10 to .16, but it also increases the average standard error of the moderation effect from .05 to .06. The ratio of the average estimate over the average standard error (analogous to an average z-statistic) increases from 2.07 to 2.56, which suggests a modest increase in power, consistent with Figure 3.4. Furthermore, when measurement error is adequately accounted for (e.g., with the latent product method) and the moderation effect is unbiased, increasing the reliability of X and Z from .70 to .90 decreases the standard error from .09 to .08, which increases the power.

3.5.3 Follow-up Monte Carlo analysis with unequal indicator reliabilities.

In these Monte Carlo analyses, we have assumed equal factor loadings of the indicators and identical indicator reliabilities. In a follow-up analysis, we also investigate the effect of unequal indicator reliabilities, to assess the appropriateness of assuming identical loadings for the indicators, as in Methods 1.1 (Means) and 2.2 (Corrected means). In addition, we investigate the effect of pairing the indicators (Method 2.3) when they have different loadings. The design varied the sample sizes from 50 to 1,500, with fixed reliabilities of X and Z at .80, effect sizes of the main and moderation effects at .20, and the correlation

between X and Z at .20. The Monte Carlo simulation with equal reliabilities generated all $\lambda = 1$ and all error variances var(ε) = .75 (Grewal et al. 2004) but unequal reliabilities, with $\lambda_{X1} = 1$, $\lambda_{X2} = 1.5$, $\lambda_{X3} = .50$, $\lambda_{Z1} = 1$, $\lambda_{Z2} = 1.5$, and $\lambda_{Z3} = .50$.

We use Methods 1.1 (Means), 1.2 (Multi-group), 2.1 (Factor scores), 2.2 (Corrected means), and 2.4 (Latent product) to estimate the moderation effect. A literature search revealed three additional indicator pairing approaches that account for unequal indicator reliabilities (Foldnes and Hagtvet 2014; Marsh et al. 2004). Method 2.3 (Product indicators: "reliability-match") uses three pairs and matches the indicators with the highest reliability from X and Z with each other, as recommended by Marsh et al. (2004). Method 2.3 (Product indicators: "reliability-compensate") uses three product indicators and combines indicators that have low reliability in X with indicators that have high reliability in Z.

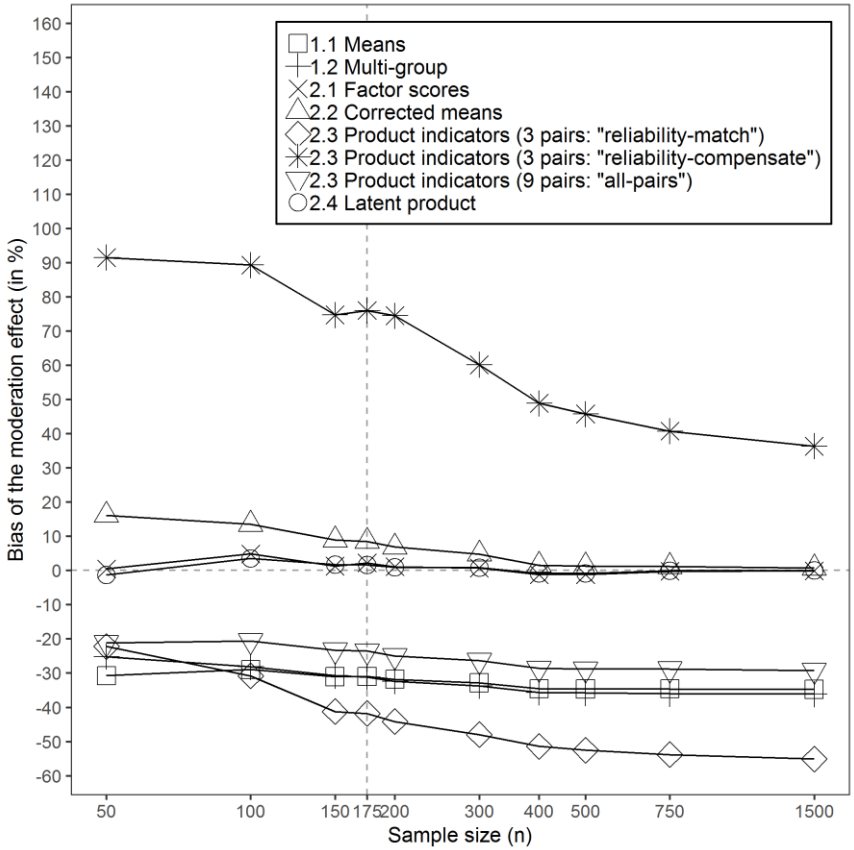
The squared loading divided by the sum of the squared loading and the error variance provides an estimate of the reliability for each indicator (Fornell and Larcker 1981). Both methods allocate indicators to pairs according to inspections of the data, which is questionable and akin to parceling methods based on indicator loadings (Foldnes and Hagtvet 2014; Little et al. 2002). Using all possible pairs of the indicators, as in Method 2.3 (Product indicators: "all-pairs"), can be useful to avoid the ambiguity of choosing them with inspection methods, as recommended by Foldnes and Hagtvet (2014).

The findings for power are consistent with the main Monte Carlo analysis. For brevity, we focus on the bias of the moderation effect in the plot on the left; the plot on the right in Figure 3.5 provides the results for the same design with equal reliabilities for comparison. The bias of the moderation effect estimated by Methods 1.1 (Means), 1.2 (Multi-group), 2.1 (Factor scores), 2.2 (Corrected means), and 2.4 (Latent product) is virtually identical to that obtained with the Monte Carlo simulation with equal indicator reliabilities.

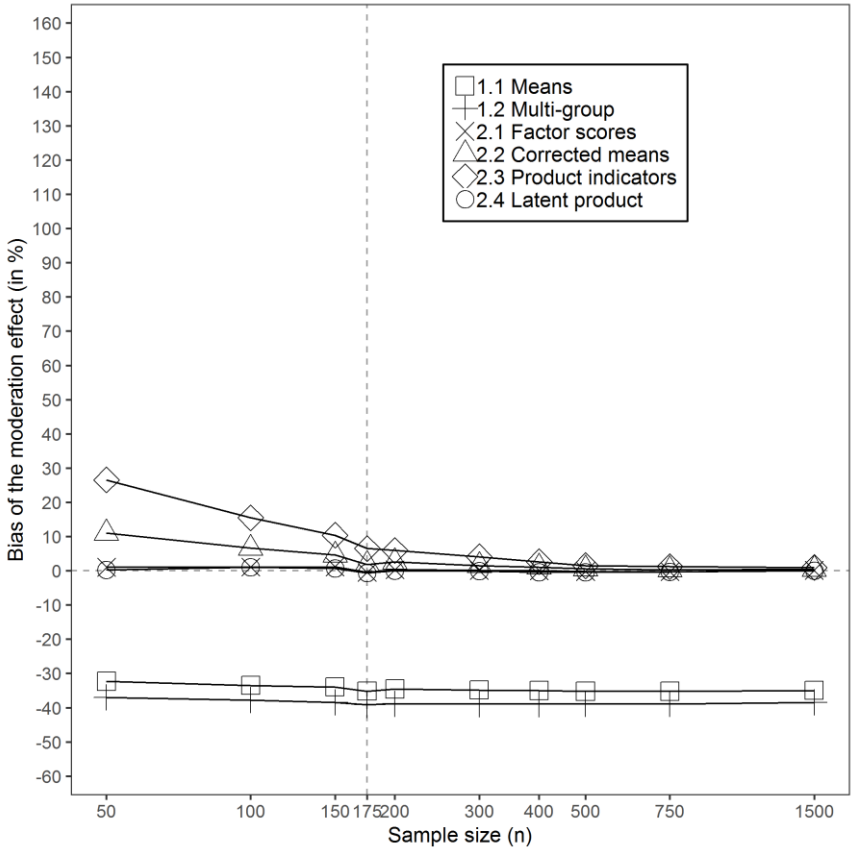
Figure 3.5

Effect of Unequal Reliabilities and Indicator Pairing on the Bias of the Moderation Effect

Panel A: Bias of the moderation effect (unequal loadings)



Panel B: Bias of the moderation effect (equal loadings)



Notes: Left plot contains results from the follow-up Monte Carlo with unequal loadings, and right plot has the results for equal loadings with the same design. Icons denote the predicted conditional means of bias of the moderation effect (in %), estimated by meta-ANOVAs with the Monte Carlo design factors and eight moderation methods as main effects, and all two-way interactions. Solid lines are interpolations between the icons. Sample size is on a log-scale. Dotted lines in both plots indicate a bias of zero (horizontal line) and a sample size of 175 (vertical line).

Method 2.3 (Product indicators) is relatively unbiased when the indicators have equal reliabilities, but all the product indicator pairings lead to severely biased moderation effects in the case of unequal reliabilities. Even for a sample size of 1,500, the "reliability-match" and "all-pairs" approaches (Foldnes and Hagtvet 2014) to construct product indicators are biased downward, by about 55% and 30%. The "reliability-compensate" approach is biased upward, by around 35%.

3.6 Discussion

We compare six methods for moderation analysis in the presence of measurement error: means, multi-group, factor scores, corrected means, product indicators, and latent product. A literature review of 504 moderation effects published in *JM* and *JMR* between 2000 and 2017 documents their usage. Monte Carlo simulations with equal and unequal indicator reliabilities provide indications of the bias and statistical power of the six methods.

The results clearly recommend against using the means or multi-group methods, both of which lead to biased estimates of true moderation effects. When the interacting variables have reliabilities of .80, which is considered good (Peterson 1994), the moderation effects estimated by the means and multi-group methods are biased downward by more than 30%. Our recommendation to account for measurement error when estimating moderation effects is not new, but it also cannot be overemphasized (Cole and Preacher 2014; Grewal et al. 2004; Pieters 2017). The vast majority of the effects in our literature review (79%) were still estimated with the means method, which can be acceptable if the reliabilities of the interacting variables are close to 1. For example, at reliabilities of .95, the bias tends to approach 10% in large samples (Equations 3.3 and 3.4). Although this is sometimes considered reasonable (Muthén and Muthén 2002), it may be exceedingly high in marketing practice when even small inefficiencies are crucial. Moreover, only about 4% of the interacting variables in the means analyses in our sample actually achieved a reliability of .95

or higher. When one or both interacting variables are categorical and reasonably can be assumed to be free of measurement error, the multi-group method can be appropriate, but all 56 cases (11%) in our sample that used the multi-group method had continuous interacting variables.

Among methods that account for measurement error, the latent product method is preferable if indicators have equal or unequal reliabilities. According to our Monte Carlo simulations, this method estimates the moderation effect with the least bias (about 1%) and highest power. It uses all the information in the raw data and does not rely on product indicators, so it is practical to use when the interacting variables have many or an unequal number of indicators. However, this method rarely has been used in existing research (about 1% of investigated effects), likely due to its high computational cost and somewhat limited accessibility, though the nlsem program in R can remedy the latter issue (Umbach et al. 2017).[5]

The corrected means and product indicators methods are unbiased in large samples (e.g., 1,500 observations) but show a moderate bias in smaller samples. Our simulations reveal small to moderate overestimation (about 4% and 12% on average) of the moderation effect at a sample size of 175. However, the product indicators method also reveals substantially biased moderation effects between 30% and 55% when the indicators have unequal reliabilities. Another limitation is the potential ambiguity associated with selecting pairs of indicators using a matched pairs approach (Foldnes and Hagtvet 2014; Marsh et al. 2004). Similarity in wording, order in a questionnaire, or random pairing are more appropriate when there is an equal number of indicators. Instead, the matched pairs approach

[5] For illustration, we generated 1,000 data sets, as in Appendix 3A, with a sample size of 175 and estimated the moderation effect in R (Umbach et al. 2017) with an Intel i7 4790 processor, running at 3.6 GHz. The output appeared almost instantly for most methods (e.g., median .06 seconds for Method 2.1 Factor scores). The estimation for Method 2.4 (Latent product) took a median of 2.76 minutes, much longer than the time for the other methods but still reasonable. The difference becomes larger when the sample size increases or multiple moderation effects are included in the model.

becomes infeasible if there are many indicators or the number of indicators for the interacting variables are unequal. Potential solutions include creating parcels or removing items from the larger scale, but such ad hoc approaches based on data inspection have questionable validity (Foldnes and Hagtvet 2014; Little et al. 2002).

The corrected means method can deal with situations in which the interacting variables have large or unequal numbers of indicators and can be applied when one or both interacting variables are single indicators. Formally, a corrected means model that uses a mean of several variables is identical to one with single-indicator measures. Single indicators typically are less reliable than measures with multiple indicators (Petrescu 2013), which lowers the power of the moderation analysis, so bigger samples are needed to detect true moderation effects with sufficient power. Moreover, determining the reliability of single-indicator measures can be challenging, because reliability estimates are not readily available from the data (Petrescu 2013; Pieters 2017)

Surprisingly, the factor scores method performs almost equally as well as the latent product method, though previous research has challenged the usefulness of factor scores in mediation and moderation analyses (Lastovicka and Thamodaran 1991; Skrondal and Laake 2001). This method appeared in about 4% of the investigated moderation analyses. According to our simulations, it is unbiased across the investigated sample sizes (50 to 1,500) and suffers only slightly lower power than the latent product method. The latent product method requires 254 observations to find a true moderation effect of .20 with 80% power, whereas the factor scores method requires 291 observations. Moreover, an advantage of the factor scores method is that it is widely available and less computationally intensive than the latent product method, such that it offers a good alternative.

We recommend a two-step procedure to implement the factor scores method. First, extract the factor scores from factor analysis. The interaction term is a product of these factor

scores. Separate confirmatory or exploratory factor analyses for each factor can be performed if factors are uncorrelated. A blockwise confirmatory factor analysis allows correlated factors (Skrondal and Laake 2001), and can account for non-random measurement errors such as common method variance (Baumgartner and Weijters 2017). The second step estimates the target moderation model with a standard path analysis or regression, depending on the application. Appendix 3C contains code to implement the factor scores method in SPSS using exploratory factor analysis. Appendix 3D has R code for the factor scores (using confirmatory factor analysis) and latent product methods.

Finally, though the recommended methods estimate unbiased moderation effects, we recommend estimating moderation effects with reliable measures and in sufficiently large samples for adequate power. The median reliability in our literature review was .86, which is good (Peterson 1994). However, the recommended sample sizes tend to be larger than what is common in practice. For illustration, we simulated 5,000 data sets for sample sizes of 100, 171, 200, 300, 400, 500, and 600, with median effect sizes (.19) and multicollinearity (.20). At a sample size of 171, the median in the literature review, the six methods achieved the following performance (percentage bias/percentage power): -25%/58% (Means), -31%/40% (Multi-group), -2%/58% (Factor scores), 1%/57% (Corrected means), 3%/54% (Product indicators), and 0%/62% (Latent product). The interpolated required sample sizes for 80% power were 298 (Means), 481 (Multi-group), 300 (Factor scores), 301 (Corrected means), 321 (Product indicators), and 276 (Latent product). However, only 35% of the moderation effects in our literature review were tested with samples of at least 275 observations, and only 15% had samples of at least 481 observations. Therefore, and considering the overwhelming prevalence of the means and multi-group methods (89% of investigated effects), it appears likely that a substantial proportion of investigated moderation effects are underestimated, underpowered, or both, to the point that even true non-null moderation effects might not have

been detected. Finding statistically significant moderation effects in such analyses might reflect false positives; not finding true non-null effects may be false negatives.

With this foundation, our study opens several avenues for further research. First, follow-up work could extend the design of the Monte Carlo analysis. We generated three indicators for X and Y, which was the most common situation in our literature review, but examining larger or unequal numbers of indicators for the main effects might be insightful. Second, further research might investigate the utility of Bayesian estimation in this context, which performs well in finite samples and facilitates the incorporation of prior information, potentially resulting in less biased estimates and moderation tests with higher power. Grewal et al. (2013) provide an application of Bayesian estimation of moderation models in marketing; Kelava and Nagengast (2012) conduct a simulation study. Third, we only investigated independent random measurement error in the indicators and did not consider variance due to common methods. Follow-up research could investigate issues with and solutions for systematic measurement error in tandem with random measurement error in moderation models.

For example, systematic measurement error might be due to common method variance (CMV) among indicators (Baumgartner and Weijters 2017). The resulting measurement error correlations may occur *within* a measure, for instance caused by similarities in wording of survey items. Using common response scales can result in error correlations *between* measures. Although it has been shown that CMV between measures underestimates moderation effects (Siemsen et al. 2010), the impact of CMV within measures (Tepper and Tepper 1993) and the combination of random measurement error with CMV (Siemsen et al. 2010) are less well understood. CMV within measures, when unaccounted for, inflates reliability estimates due to increased correlations between indicators, resulting in correlation estimates that are not adequately disattenuated. CMV between Y and X can inflate

correlations. Yet, the joint effects of measurement error and CMV within and between measures are likely to be more complicated due to the reliability of the product term being affected by both the reliability of the components and their correlation. Monte Carlo simulations are therefore warranted to better understand the effects of random measurement error as well as CMV within and between measures on moderation effects. Simulations in this chapter can be extended by adding a multi-indicator dependent variable and different patterns of CMV within and between measures (e.g., within Y, within X, between X and Y, between X and Z). Moderation methods with measurement models that do not account for CMV, like those in this chapter, can then be compared to measurement models that do, such as those reported in Baumgartner and Weijters (2017).

In sum, the latent product method to estimate moderation effects performed best in the present simulations. It is unbiased and offers the highest power. A surprisingly poorly performing method is the product indicators method. Although it relies on confirmatory factor analysis and is commonly proposed in the methods literature, our Monte Carlo simulations reveal that it is substantially biased in small samples (e.g., 175 observations). A surprisingly good alternative is the factor scores method. As the latent product method, it also results in unbiased estimates, with only slightly lower power, but it is easier to implement and widely available. In this light, it is hard to justify the continued use of the means and multi-group methods when indicators of the interacting constructs in moderation analysis are measured with measurement error.

Appendix of Chapter 3

Appendix 3A: Hypothetical data.

Table A3.1
Hypothetical Data Input for the Six Methods for Moderation Analysis

Panel A: Correlation matrix of hypothetical data

	Y	v_{X1}	v_{X2}	v_{X3}	v_{Z1}	v_{Z2}	v_{Z3}
Y	1						
v_{X1}	.181	1					
v_{X2}	.192	.561	1				
v_{X3}	.184	.571	.556	1			
v_{Z1}	.177	.106	.085	.120	1		
v_{Z2}	.155	.117	.085	.112	.573	1	
v_{Z3}	.136	.159	.135	.123	.577	.572	1

Panel B: Correlation matrix for Method 1.1 Means

	Y	$\overline{X}$	$\overline{Z}$	$\overline{X}\overline{Z}$
Y	1			
$\overline{X}$	.220	1		
$\overline{Z}$	.185	.163	1	
$\overline{X}\overline{Z}$	.152	-.072	.006	1

Panel C: Correlation matrix for Method 2.1 Factor scores

	Y	$\hat{X}$	$\hat{Z}$	$\hat{X}\hat{Z}$
Y	1			
$\hat{X}$	.226	1		
$\hat{Z}$	.192	.246	1	
$\hat{X}\hat{Z}$	.156	-.069	-.000	1

Panel D: Correlation matrices for Method 1.2 Multi-group

Group 1	Y^1	v^1_{X1}	v^1_{X2}	v^1_{X3}
Y^1	1			
v^1_{X1}	.040	1		
v^1_{X2}	.037	.586	1	
v^1_{X3}	.081	.597	.576	1

Group 2	Y^2	v^2_{X1}	v^2_{X2}	v^2_{X3}
v^2_{X1}	1			
v^2_{X2}	.277	1		
v^2_{X3}	.315	.524	1	
v^2_{X1}	.255	.531	.525	1

Panel E: Correlation matrix and additional information for Method 2.2 Corrected means

	Y	v_{X1}	v_{X2}	v_{X3}	v_{Z1}	v_{Z2}	v_{Z3}	$\overline{X}\overline{Z}$
Y	1							
v_{X1}	.181	1						
v_{X2}	.192	.561	1					
v_{X3}	.184	.571	.556	1				
v_{Z1}	.177	.106	.085	.120	1			
v_{Z2}	.155	.117	.085	.112	.573	1		
v_{Z3}	.136	.159	.135	.123	.577	.572	1	
$\overline{X}\overline{Z}$	.152	-.034	-.080	-.068	.035	.008	-.029	1

r_{XX} = .795
r_{ZZ} = .801
$r_{X,Z}$ = .163
σ^2_{XZ} = 1.610
$r_{XZ,XZ}$ = .646

Table A3.1 (CONTINUED)

Panel F: Correlation matrix for Method 2.3 Product indicators

	Y	v_{X1}	v_{X2}	v_{X3}	v_{Z1}	v_{Z2}	v_{Z3}	$v_{X1}\times v_{Z1}$	$v_{X2}\times v_{Z2}$	$v_{X3}\times v_{Z3}$
Y	1									
v_{X1}	.181	1								
v_{X2}	.192	.561	1							
v_{X3}	.184	.571	.556	1						
v_{Z1}	.177	.106	.085	.120	1					
v_{Z2}	.155	.117	.085	.112	.573	1				
v_{Z3}	.136	.159	.135	.123	.577	.572	1			
$v_{X1}\times v_{Z1}$	.139	-.002	-.030	-.007	.051	.019	-.005	1		
$v_{X2}\times v_{Z2}$	.099	-.065	-.091	-.079	.033	.021	-.033	.357	1	
$v_{X3}\times v_{Z3}$	.088	-.054	-.079	-.090	-.002	-.022	-.005	.397	.372	1

Panel G: Raw data for Method 2.4 Latent product

	Y	v_{X1}	v_{X2}	v_{X3}	v_{Z1}	v_{Z2}	v_{Z3}
1	-1.289	-.506	-1.007	-.887	-1.677	-1.481	-1.957
2	-.012	.469	-.680	.325	.228	.823	.092
3	.177	2.080	1.645	1.017	-1.042	-.639	-.934
…	…	…	…	…	…	…	…
n	.316	-.400	.159	.210	-1.597	-.710	-.958

Notes: Panel A is a correlation matrix of hypothetical data generated with Equations 3.1 and 3.2, reliabilities of X and Z of .80, $\beta_1 = \beta_2 = \beta_3 = .20$, and the correlation between X and Z is .20. Panels B to G represent data input derived from Panel A for the six moderation methods.

Appendix 3B: Bias and power of the moderation effect in the Monte Carlo.

For each cell, we calculated bias and power across the data sets that converged for all methods. A robust method that converges on a data set while other methods do not converge would otherwise be penalized, in that the resulting extreme estimates could lead to high bias.

Let k ϵ (1, 2, …, K) denote the index for the Monte Carlo replications where K is the number of replications that converged for all methods in that cell. The bias is:

$$\text{Bias}_{\beta_3} = 100 \times \frac{\left(\frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_{3,k} - \beta_{3,}\right)}{\beta_3}, \tag{A3.1}$$

where $\hat{\beta}_{3,k}$ is the estimated moderation effect for data set k, and β_3 is its population value. The power is:

$$\text{Power}_{\hat{\beta}_3} = \frac{1}{K}\sum_{k=1}^{K} I(\cdot), \tag{A3.2}$$

with

$$I(\cdot) = \begin{cases} 1 \text{ if } \left|\frac{\hat{\beta}_{3,k}}{SE(\hat{\beta}_{3,k})}\right| > 1.96 \\ 0 \text{ otherwise} \end{cases}, \tag{A3.3}$$

where $SE(\hat{\beta}_{3,k})$ is the standard error of the moderation effect, and 1.96 is the critical value.

The multi-group method does not estimate β_3, and we calculated the estimated moderation effect for each replication by calculating the z-statistic:

$$z - \text{statistic}_{\text{multi-group,k}} = \frac{\hat{\beta}_{1,k}^{g=2} - \hat{\beta}_{1,k}^{g=1}}{\sqrt{SE\left(\hat{\beta}_{1,k}^{g=1}\right)^2 + SE\left(\hat{\beta}_{1,k}^{g=2}\right)^2}}, \tag{A3.4}$$

where g = 1 denotes group membership below the median of z, and g = 2 is above the median of z. The estimated moderation effect size is then (Rosenthal and DiMatteo 2001):

$$\hat{\beta}_{3,k} = \frac{z - \text{statistic}_{\text{multi-group,k}}}{\sqrt{n}}, \tag{A3.5}$$

where n is the sample size**.**

Appendix 3C: SPSS code for the factor scores method.

```
* Example SPSS Code for Method 2.1 (Factor scores).
* Y is the dependent variable.
* vx1-vx3 and vz1-vz3 are indicators for X and Z.

* Step 1: Perform factor analysis and estimate factor scores.

FACTOR /* Exploratory factor analysis for X.
 /VARIABLES vx1 vx2 vx3 /* three indicators for X.
 /CRITERIA FACTORS(1) /* extract one factor.
 /EXTRACTION ML /* use maximum likelihood.
 /ROTATION NOROTATE /* no rotation.
 /SAVE REG(ALL, FX). /* save factor score of X as "FX1".

FACTOR /* Exploratory factor analysis for Z.
 /VARIABLES vz1 vz2 vz3 /* three indicators for Z.
 /CRITERIA FACTORS(1) /* extract one factor.
 /EXTRACTION ML /* use maximum likelihood.
 /ROTATION NOROTATE /* no rotation.
 /SAVE REG(ALL, FZ). /* save factor score of Z as "FZ1".

COMPUTE FXZ = FX1 * FZ1. /* Compute XZ.
EXECUTE.

* Step 2: Estimate the target moderation model.

REGRESSION /* Linear regression of Y on X, Z and XZ.
 /DEPENDENT Y
 /METHOD=ENTER FX1, FZ1, FXZ.
```

Appendix 3D: R code for the factor scores and latent product methods.

```
# Example R code for Method 2.1 (Factor scores)
# Data is in object 'data'
# Y is the dependent variable
# vx1-vx3 and vz1-vz3 are indicators for X and Z

library(lavaan) # package for latent variable analysis: Rosseel, Yves
(2012), "lavaan: An R Package for Structural Equation Modeling," Journal of
Statistical Software, 48 (2), 1-36.

# Step 1: Perform factor analysis and estimate factor scores

cfamodel <- ' # confirmatory factor analysis (CFA) for X and Z
X =~ vx1 + vx2 + vx3
Z =~ vz1 + vz2 + vz3 '

cfafit <- cfa(model = cfamodel, data = data) # fit the CFA
fscores <- lavPredict(cfafit, method = "regression") # estimate factor
scores
data <- cbind(data, fscores) # add factor scores to the data
data$XZ <- data$X*data$Z # compute the product of factor scores XZ

# Step 2: Estimate the target moderation model

model2.1 <- ' Y ~ X + Z + XZ ' # path analysis of Y on X, Z, and XZ

summary(sem(model = model2.1, data = data)) # estimate the path analysis

# Example R code for Method 2.4 (Latent product)
# Data is in object 'data'

library(nlsem) # package for latent product method: Umbach, Nora, Katharina
Naumann, Holger Brandt, and Augustin Kelava (2017), "Fitting Nonlinear
Structural Equation Models in R with Package nlsem," Journal of Statistical
Software, 77 (1), 1-20.

colnames(data) = c("x1","x2","x3","x4","x5","x6","y1") # rename indicators

model2.4 <- specify_sem(num.x = 6, # indicators for exogenous variables
 num.y = 1, # indicators for endogenous variable,
 num.xi = 2, # latent exogenous variables
 num.eta = 1, # endogenous variables
 xi= "x1-x3, x4-x6", # 3 indicators for X and 3 for Z
 eta = 'y1', # 1 indicator for Y
 interaction = "eta1 ~ xi1:xi2") # the interaction

set.seed(51585) # set a seed for reproducibility
start <- runif(count_free_parameters(model2.4)) # set starting values

fit2.4 <- em(model2.4, data, start, verbose = TRUE) # estimate the model
summary(fit2.4) # print the results in the console
```

Chapter 4 – Discriminant Validity for Meaningful Process Analysis in Marketing Research

4.1 Introduction

Construct validity is essential for meaningful theory testing in marketing research. It refers to the correspondence between latent constructs and their measures. This chapter examines discriminant validity, which is one of the preconditions for construct validity (Peter 1981). Discriminant validity is achieved when measures of theoretically distinct constructs are empirically distinct (Campbell and Fiske 1959). Without discriminant validity between constructs in the hypothesized theory there can be no meaningful theory test (Strauss and Smith 2009). Establishing discriminant validity prevents redundant constructs and the proliferation of increasingly fine-grained constructs which might be semantically distinct but which cannot be empirically distinguished from each other (Shaffer et al. 2016). Without discriminant validity, constructs with hypothesized relationships between them may in fact be indistinguishable measures of a single underlying construct. Distinctiveness of measures of constructs is thus also a necessary condition for meaningful process analyses. Lack of discriminant validity between successive stages in process models thwarts making causal inferences (Pieters 2017; Spencer et al. 2005). Moreover, theoretically distinct facets of a multi-faceted construct can lack discriminant validity which prevents making the nuanced inferences that the multi-faceted construct intended to provide. Statistically, high correlations between measures of constructs can lead to estimation issues such as excessive multicollinearity, model non-convergence and to inaccurate estimates of parameters and standard errors if the model converges. This can result in substantial Type II errors and a failure to detect true effects (Grewal et al. 2004).

Despite its pivotal role in theory testing, marketing research has devoted remarkably little attention to discriminant validity, with large differences between study methodologies. Specifically, Hulland et al. (2018) found in a review of 202 survey-based studies published in

the *Journal of the Academy of Marketing Science* between 2006 and 2015 that about 78% of survey-based studies did report one or more tests of discriminant validity. Likewise, Voorhees et al. (2016) found in an ambitious review of 1,931 articles published between 2008 and 2012 in seven leading marketing journals that 77.8% of survey-based studies reported on discriminant validity. However, and importantly, only 3.5% of studies that used experiments did, and experiments make up the majority of studies in consumer research (Peighambari et al. 2016). Similarly, Pieters (2017) found that only three out of 166 mediation analyses (less than 2%) in 86 articles using experiments published in the *Journal of Consumer Research* between 2014-2016 assessed discriminant validity between mediator and outcome. Also, Voorhees et al. (2016) found that less than one percent of studies using secondary data addressed discriminant validity. Thus, while a substantial percentage of survey-based studies report on discriminant validity, it appears to be almost ignored in other areas of marketing and research.

One reason for the overall scant attention to discriminant validity could be the belief among researchers that a theoretical rationale for the distinctiveness of measures of constructs is a sufficient condition for discriminant validity. The analyst then assumes discriminant validity without further testing for it. Another reason may be the belief that establishing reliability of measures is sufficient for construct validity. In fact, Shook et al. (2004) found that 61% out of 92 studies that used structural equation modeling and were published in nine leading strategic management journals between 1985 to 2002 reported on reliability of measures, whereas only 40% reported on the discriminant validity. Reliability of measures and validity of constructs are indeed related because discriminant validity is less likely when the reliability of measures of constructs is low (Pieters 2017). Yet, two measures that have perfect reliability can still be perfectly correlated which would make them statistically indistinguishable. Furthermore, the belief may exist that discriminant validity is only relevant

for measures with multiple indicators but not for single-indicators, or that the reliability of single-indicators cannot be assessed. This may be reflected in the result that less than one percent of studies that used secondary data, which presumably mostly use single-indicators, assessed discriminant validity (Voorhees et al. 2016). In practical applications, two single-indicators of presumably different constructs can be perfectly correlated, despite their theoretical distinctiveness. Finally, the belief may exist that establishing construct validity requires specialized statistical analyses that are not commonly available in conventional statistical software packages (MacKenzie 2001; Steenkamp and van Trijp 1991). Irrespective of the reasons, discriminant validity is given very little attention in marketing research.

Even more, discussions of discriminant validity in marketing to date have exclusively focused on bivariate discriminant validity. Bivariate discriminant validity captures the empirical distinctiveness within each pair of measures of constructs (Anderson and Gerbing 1988; Fornell and Larcker 1981; Franke and Sarstedt 2018; Voorhees et al. 2016). It sets out to detect situations where pairs of measures cannot be empirically distinguished. However, two theoretically distinct measures of constructs may express pairwise discriminant validity with respect to each other but can jointly perfectly account for a measure of another construct. For example, a first measure may be correlated .70 with a second measure and .70 with a third one, and the second and third may be uncorrelated with each other. The second and third measures then almost perfectly account for the first one (shared variance = 98%), whereas it only shared about half of its variance pairwise ($.70^2 \times 100\% = 49\%$). For example, an outcome can be perfectly accounted for by two parallel mediators which makes the measures in the mediation and outcome stage indistinguishable. Lack of discriminant validity can also lead to empirical issues such as model non-convergence and Type II errors due to multicollinearity (Grewal et al. 2004).

The objective of this chapter is threefold. First, it extends conventional bivariate discriminant validity criteria with a new multivariate discriminant validity criterion. The criterion accounts for all correlations between measures of constructs in a set instead of assessing pairs of measures. This is important because each of the pairs of measures could empirically pass the bivariate criterion, whereas a pair can jointly fully account for the variance in a third one such that multivariate discriminant validity is not supported. By presenting this approach, we answer a recent call by Franke and Sarstedt (2018, p. 442) to develop criteria for discriminant validity that consider networks of constructs rather than pairs of constructs. Our multivariate approach follows up on the idea that "'[l]earning more about' a theoretical construct is a matter of elaborating the nomological network in which it occurs…" (Cronbach and Meehl 1955, p. 12). Constructs derive their meaning from their position in nomological networks of related constructs, and construct validity should be assessed using a network of associations in which the construct occurs (Cronbach and Meehl 1955). The proposed discriminant validity criterion is relevant for theories with more than two constructs, or multidimensional constructs which are abound in marketing research such as need for uniqueness (Tian et al. 2001), materialism (Richins and Dawson 1992), and market-orientation (Narver and Slater 1990). Establishing multivariate discriminant validity is also necessary if measures of multiple constructs jointly account for a dependent variable, such as in simple mediation and multiple mediation with parallel (Müller-Stewens et al. 2017) or sequential mediators (Bellezza et al. 2017), as well as theories with multiple dependent variables (Auh et al. 2019).

Second, this chapter provides a quantitative literature review of multiple mediation models in marketing and research to illustrate common practice with respect to discriminant validity assessment in an important theory testing domain. Among 23 studies that were recently published in the *Journal of Marketing Research*, *Journal of Marketing* and *Journal*

of Consumer Research, we find that only 13 reported on discriminant validity. Nevertheless, high correlations (up to .92) occur between the measures of the constructs which puts them at risk of lacking discriminant validity. Four case studies demonstrate situations (such as high correlations, low reliabilities, small sample size) that are at a particular risk of lack of discriminant validity. The reanalyses cast doubt on the validity of the purported multiple mediation theories.

Third, this chapter aims to make testing for bivariate and multivariate discriminant validity more accessible to analysts. It offers an online application which facilitates establishing bivariate and multivariate discriminant validity, using summary statistics data (SSD) only. SSD are a compact, aggregate, form of raw data and can be readily included in reports (Pieters 2017). SSD for linear (regression, ANOVA and structural equation) models typically consist of a correlation matrix of all measures and treatments (in case of an experiment), the estimated reliabilities of measures with multiple indicators, and the sample size. Using SSD is useful in situations where the raw data are not available, such as during study planning, study evaluation, or meta-analysis. For this purpose, the online tool facilitates the use of discriminant validity methods requiring only SSD. The online application is available at: https://github.com/constantpieters/dv. It includes case studies that can be readily used.

4.2 Discriminant Validity

Discriminant validity has both theoretical and empirical facets (Shaffer et al. 2016). It refers to the distinctiveness of constructs as well as their measures. Adding a new or existing construct to a theoretical (process) model does not only specify the theoretical pathways that relate the focal construct with other constructs (nomological validity). It also puts forward an, often implicit, theory that the focal construct is discriminant valid with respect to other constructs (Harter and Schmidt 2008), even from constructs that might be related but are not

included in the focal model. On the theoretical level, constructs in theories must be then be defined and demarcated to demonstrate their theoretical distinctiveness with respect to other constructs. According to Le et al. (2010, p. 113), "[b]ecause of the conceptual/theoretical fluency of researchers, [the theoretical discriminant validity facet] is essentially a weak one and is usually easily met." Yet, the distinction between theoretical and empirical discriminant validity is an important one. Although we may be able to conceptualize an endless amount of fine-grained and semantically different constructs, their measures could not always be empirically distinguishable, which is a threat to construct validity. Importantly, empirical discriminant validity is then achieved when measures of theoretically distinct constructs are empirically distinct (Campbell and Fiske 1959). It is an empirical validation of the often implicit construct distinctiveness theory, and contributes to construct validity (Peter 1981). This chapter focuses on such empirical discriminant validity (hereinafter, "discriminant validity" for brevity).

4.2.1 Discriminant validity within and between model stages.

Figure 4.1 shows several hypothetical models to conceptualize and illustrate discriminant validity.[6] Each model distinguishes the theoretical (graphical representation of hypothesized relationships between constructs) from the statistical (equations that specify and identify relationships between measures of constructs) model posited by the theory, to be used below. Models 1.1 (bivariate regression), 1.2 (multivariate regression), and 1.3 (multiple regression) are non-process models with direct and non-moderated relationships between X and Y constructs. They all have two stages: X and Y.

Models 2.1 (basic mediation), 2.2 (sequential multiple mediation), 2.3 (parallel multiple mediation), and 3 (general process model with multiple inputs, mediators, and

[6] We omit interactions (moderation) between variables in our examples for ease of exposition. However, the examples, framework and discussion readily generalize to include moderation and interaction variables.

Figure 4.1
Hypothetical Theoretical and Statistical Models

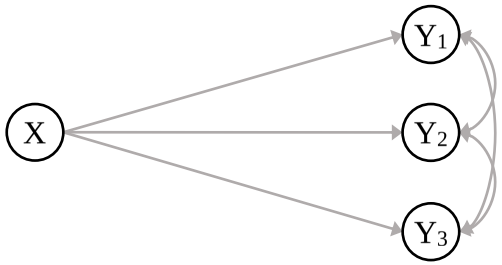
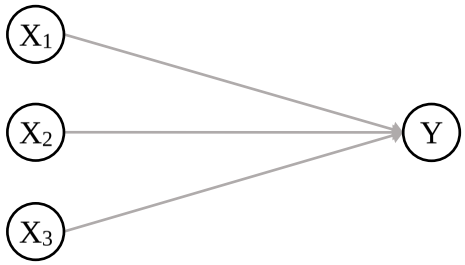
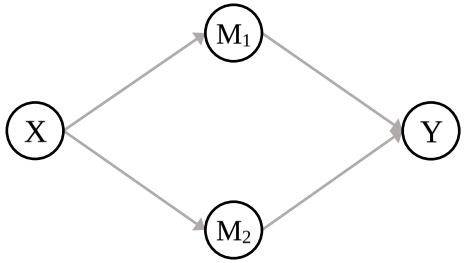
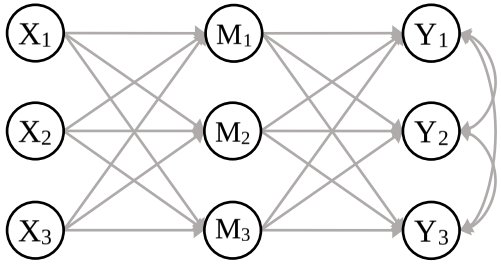
Model 1.1: Bivariate regression	Model 1.2: Multivariate regression	Model 1.3: Multiple regression
Theoretical model	*Theoretical model*	*Theoretical model*
		
Statistical model $Y = cX$	*Statistical model* $Y_1 = c_1X$; $Y_2 = c_2X$; $Y_3 = c_3X$	*Statistical model* $Y = c_1X_1 + c_2X_2 + c_3X_3$
Discriminant validity Not applicable	*Discriminant validity* Not applicable	*Discriminant validity* Within X stage
Model 2.1: Basic mediation	**Model 2.2: Sequential multiple mediation**	**Model 2.3: Parallel multiple mediation**
Theoretical model	*Theoretical model*	*Theoretical model*
		
Statistical model $M = aX$ $Y = bM + cpX$ Indirect effect X on Y = $a \times b$	*Statistical model* $M_1 = a_1X$; $M_2 = dM_1 + a_2X$ $Y = b_1M_1 + b_2M_2 + cpX$ Indirect effect X on Y = $a_1 \times b_1 + a_2 \times b_2 + a_1 \times d \times b_2$	*Statistical model* $M_1 = a_1X$; $M_2 = a_2X$ $Y = b_1M_1 + b_2M_2 + cpX$ Indirect effect X on Y = $a_1 \times b_1 + a_2 \times b_2$
Discriminant validity Between X, M, and Y stages (if indirect effect is hypothesized and estimated)	*Discriminant validity* Within & between X, M & Y stages (if indirect effect is hypothesized and estimated)	*Discriminant validity* Within & between X, M & Y stages (if indirect effect is hypothesized and estimated)

Figure 4.1 (CONTINUED)

Model 3: General process model with multiple inputs, mediators, and outcomes

Theoretical model



Statistical model

$M_1 = a_1X_1 + a_2X_2 + a_3X_3$; $M_2 = a_4X_1 + a_5X_2 + a_6X_3$; $M_3 = a_7X_1 + a_8X_2 + a_9X_3$

$Y_1 = b_1M_1 + b_2M_2 + b_3M_3 + cp_1X_1 + cp_2X_2 + cp_3X_3$; $Y_2 = b_4M_1 + b_5M_2 + b_6M_3 + cp_4X_1 + cp_5X_2 + cp_6X_3$; $Y_3 = b_7M_1 + b_8M_2 + b_9M_3 + cp_7X_1 + cp_8X_2 + cp_9X_3$

Indirect effect X on $Y_1 = a_1{\times}b_1 + a_2{\times}b_1 + a_3{\times}b_1 + a_4{\times}b_2 + a_5{\times}b_2 + a_6{\times}b_2 + a_7{\times}b_3 + a_8{\times}b_3 + a_9{\times}b_3$

Indirect effect X on $Y_2 = a_1{\times}b_4 + a_2{\times}b_4 + a_3{\times}b_4 + a_4{\times}b_5 + a_5{\times}b_5 + a_6{\times}b_5 + a_7{\times}b_6 + a_8{\times}b_6 + a_9{\times}b_6$

Indirect effect X on $Y_3 = a_1{\times}b_7 + a_2{\times}b_7 + a_3{\times}b_7 + a_4{\times}b_8 + a_5{\times}b_8 + a_6{\times}b_8 + a_7{\times}b_9 + a_8{\times}b_9 + a_9{\times}b_9$

Discriminant validity

Within X & M stages, between X & M & Y stages

(if indirect effects are hypothesized and estimated)

Notes: Circles are constructs and straight arrows are causal relationships between constructs. X refers to input, M to mediator (throughput) and Y to outcome. Curved arrows are residual covariances. Indicators (measures) and residuals are omitted for exposition. Covariances between inputs and residual covariances between (parallel) mediators are omitted from theoretical and statistical models for brevity. Model 3 readily generalizes to sequential mediation or a combination of parallel and sequential mediation by adding “d-paths” such as in Model 2.2. The framework readily generalizes to moderation with interaction effects.

outcomes) are process models with mediation. They have three stages, X, M and Y, determined by their theoretical status in the model: input (X), throughput (M) or output (Y).

We propose two guidelines to inform meaningful discriminant validity assessment. One for any model and a second for mediation models that decompose total effects into indirect and conditional direct effects (Pieters 2017). First, discriminant validity in models is required *within* X or M stages that contain more than one construct (e.g., X_1, X_2, and X_3 or M_1, M_2 and M_3). In other words, measures of constructs within a stage that enter on the right side of an equation in a statistical model need to be distinct. Discriminant validity within stages is essential to identify and quantify theoretically distinct effects. For instance, if X_1 and X_2 are not empirically distinct from X_3, their theoretical pathways to Y cannot be empirically separated in the multiple regression represented by Model 1.3 in Figure 4.1. Managers might not be able to manipulate X_1, X_2 or X_3 separately to influence Y. In fact, a single X measure that regresses on Y (Model 1.1) might then best represent the effects of the three measures X_1 to X_3. As an example, if processing motivation (X_1), ability (X_2) and opportunity (X_3) of advertisements are not discriminant valid, their separate effects on levels of brand processing (Y) cannot be distinguished (Maclnnis et al. 1991). Similarly, for meaningful effects of M_1 and M_2 on Y, the M_1 and M_2 measures must be distinct (Model 2.2). Note that for measures of outputs Y, discriminant validity does not have to be attained within the Y-stage because such outputs by definition do not appear in equations of other outputs.

Second, discriminant validity should be determined *between* stages to decompose total effects into indirect and direct effects in mediation models (Pieters 2017; Spencer et al. 2005). In other words, all measures of constructs that are on the right side of a focal Y-equation of a mediation model should be distinct among each other *and* with the focal Y. As an example, if X and M, X and Y, or M and Y in a basic mediation model (Model 2.1 in Figure 4.1) are not discriminant valid, the a, b and cp paths cannot be meaningfully separated

and the model might best reduce to the bivariate regression in Model 1.1. Similarly, if M_2 and Y are not distinct in the serial mediation Model 2.2, the b2 term in the indirect effect cannot be meaningfully estimated, the d and b1 paths as well as the a2 and cp paths cannot be separated, yielding a non-meaningful expression a1×b1 + a2×b2 + a1×d×b2 of the indirect effect. Then, the serial mediation model might best be reduced to the basic mediation Model 2.1. Note that this condition does not preclude partial or full mediation from taking place. The distinction between partial and full mediation is about the proportion of the total effect being mediated, discriminant validity is about the proportion of the variance in a measure being explained by the others, as detailed below.

An important condition for between stage discriminant validity is the expectation of indirect effects. Models with throughput variables M that do not hypothesize mediation or indirect effects do not require discriminant validity between stages. For instance, a restricted Model 2.3 that does not estimate cp (or fixes it to zero), the conditional direct effect of X on Y, does not investigate mediation and cannot adequately estimate an indirect effect. It is equivalent to estimating the M and Y equations separately without conditional direct effects. Hence, discriminant validity between stages does not have to be attained, although the discriminant validity requirements within stages remain to properly separate the b-paths. Moreover, the two guidelines also imply that bivariate or multiple regressions depicted in Models 1.1 and 1.3 do not require discriminant validity between stages. In such models, moderately fitting models for Y would be preferred over well-fitting models for Y. Although model fit on itself might not be an indicator of the quality of a theory (Roberts and Pashler 2000), it might inform predictions, and can be managerially relevant.

These principles generalize to larger and more complex models. For instance, the process Model 3 in Figure 4.1 requires discriminant validity between stages, for instance of Y_1 with respect to all X and M measures to meaningfully identify indirect effects. Of course,

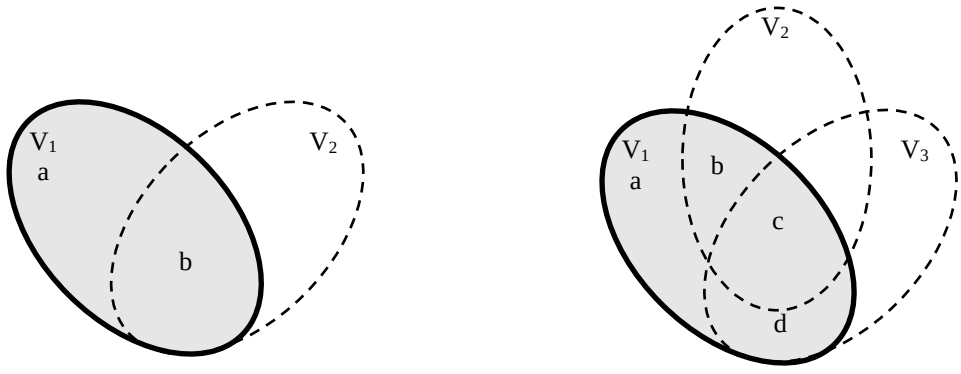
although evidence for discriminant validity can be found in a model, there can still be considerable levels of multicollinearity, an issue closely related to that of discriminant validity (Grewal et al. 2004). Multicollinearity refers to excessive correlations between predictors on the right side of an equation (Greene 2008), that might be within or between stages depending on the full statistical model. Even moderate levels of multicollinearity may result in empirical issues such as Type II errors in small samples, when reliability is low, or when overall model fit is poor (Grewal et al. 2004; Kalnins 2018; Mason and Perreault 1991). A follow-up analysis in Chapter 5 returns to multicollinearity and its implications.

4.2.2 Bivariate (BDV) and multivariate discriminant validity (MDV).

Discriminant validity can be assessed for pairs of measures (bivariate) or in sets of three or more measures of constructs (multivariate). Bivariate discriminant validity (BDV) assesses the associations of pairs of measures, regardless of the set size. Yet, measures of two constructs might attain BDV with respect to each other, and with respect to a third measure. The third measure may then not be distinct from a combination of the other two. Multivariate discriminant validity (MDV) sets out to identify such situations and is assessed for a focal

Figure 4.2
Tulipograms of Bivariate (BDV) and Multivariate Discriminant Validity (MDV)

Panel A: Two Measures of Constructs (Bivariate) Panel B: Three Measures of Constructs (Multivariate)



Notes: Petals (ellipses) in the Tulipograms are measures of constructs. The focal measure is V_1 (colored grey with bold solid outline). Segments within V_1 are variance contributions. In the left plot, segment a is the coefficient of alienation and b refers to the coefficient of determination. In the right plot, a is the coefficient of alienation and b & d are the unique variance contributions of respectively V_2 and V_3 to V_1, c is the joint variance contribution of V_2 and V_3 to V_1, and segments b to d sum to the coefficient of determination of V_1.

measure with respect to all other measures in a set, taking into account all associations between the measures in that set.

We illustrate the distinction of BDV and MDV with Tulipograms. Figure 4.2 contains such Tulipograms, which derive their name from the resemblance of a tulip. The Tulipogram is inspired by a Venn diagram which can be used to express associations between measures of constructs (Cohen et al. 2003). The ellipses, petals of the tulip, represent measures of constructs and the shaded segments in the ellipses are variance contributions to the focal measure V_1. We use ellipses because "[b]eyond three terms circles fail us, since we cannot draw a fourth circle which shall intersect three others in the way required" (Venn 1880, p. 6).

Panel A in Figure 4.2 presents a bivariate variance decomposition. It uses "V" notation for measures of constructs to abstract from model stages. The total variance of the focal measure of construct V_1 is decomposed into proportions a and b. Segment a represents the coefficient of alienation: the proportion of unique variance in V_1. Segment b is the coefficient of determination: the proportion of variance of V_1 that is shared with V_2 (Cohen et al. 2003).

Panel B in Figure 4.2 contains a graphical representation of a variance decomposition of V_1 with two other measures of constructs, V_2 and V_3. Segment a is the coefficient of alienation. The coefficient of determination is segments b + c + d. Segment b represents the *unique* variance contribution of V_2 to V_1, and segment d that of V_3 to V_1. Segment c is the joint contribution of V_2 and V_3 to the variance of V_1. It can take negative values, for example in case of suppression (Cohen et al. 2003; Friedman and Wall 2005).

Now that we have introduced BDV and MDV, we present and extend empirical BDV and MDV criteria and thresholds. We then discuss the impact of measurement error on BDV and MDV and present estimated power curves of MDV.

4.3 Empirical Assessment of Discriminant Validity.

4.3.1 Bivariate discriminant validity (BDV).

In the bivariate case, squaring the correlation between two measures is an estimate of the coefficient of determination of both measures. There have been various proposals for discriminant validity criteria which focus on this correlation (Anderson and Gerbing 1988; Fornell and Larcker 1981; Henseler et al. 2015; Jöreskog 1971). A general BDV criterion is then:

$$|r_{V1.V2}| < T, \tag{4.1}$$

where $r_{V1,V2}$ is the correlation between V_1 and V_2, and T is some threshold.

Table 4.1 shows an overview of potential thresholds. It distinguishes correlation-based thresholds and reliability-based thresholds, and the remainder of this chapter uses both. Correlation-based thresholds use fixed values of correlations. A common correlation-based threshold is unity, T = 1 (Anderson and Gerbing 1988; Fornell and Larcker 1981; Voorhees et al. 2016). The criterion then establishes whether the coefficient of determination is smaller than one (i.e., measures of both constructs are distinct), or analogously, whether the coefficient of alienation is larger than zero (i.e., there is unique variance in a measure). Other ad hoc correlation-based thresholds than T = 1 have been proposed with little theoretical foundation. For example, Bagozzi and Yi (1988) suggested .95, and Henseler et al. (2015) proposed to use .90, .95 or .85 as threshold. A stricter threshold is .71 (MacKenzie et al. 2011). Meeting the discriminant validity criterion with this threshold suggests that the shared variance of a measure with another construct is less than 50% ($.71^2 \times 100\% \approx 50\%$ shared variance between measures of constructs), i.e., there is more than half unique variance in a measure.

Reliability-based thresholds use reliability information as thresholds for discriminant validity. A reliability-based criterion assesses whether the correlation between measures of

constructs (between construct correlation) is smaller than the reliability of the measure which is the correlation of one or more measures with the construct they purport to capture (Peter 1981), a within construct correlation. Using a reliability-based threshold therefore uses correlations between and within constructs, and the correlation-based threshold uses only the correlation between constructs. Fornell and Larcker (1981) proposed a similar criterion in that the variance shared (squared correlation) between two constructs has to be smaller than the average variance extracted, which is an estimator of reliability (Baumgartner and Homburg 1996) based on shared variances between indicators of the same construct.

It is important to note that the discriminant validity criterion in Equation 4.1 can be assessed with statistical tests and heuristics, providing directional evidence that the correlation is smaller than the threshold (Franke and Sarstedt 2018). Whereas using heuristics is accessible, it does not account for the statistical uncertainty in the correlation estimate and the threshold, if a reliability-based criterion is used. A recent simulation study advocates the use of statistical discriminant validity tests by showing that using heuristics has lower power and produces more false positives (Franke and Sarstedt 2018).

Table 4.1
Discriminant Validity Thresholds

Threshold (T)	Foundation	Reference(s)
Correlation-based: discriminant validity is assessed using only between construct correlations		
T = 1	Theoretical: target measure has unique variance	Anderson and Gerbing (1988)
T = .95	Ad hoc	Henseler et al. (2015)
T = .90	Ad hoc	Henseler et al. (2015)
T = .85	Ad hoc	Voorhees et al. (2016)
T = .71	Theoretical: proportion of unique variance in a target measure > 50%	MacKenzie et al. (2011)
Reliability-based: discriminant validity is assessed using between and within construct correlations		
$T = r_{ViVi}$	Theoretical: target measure has a higher correlation with the construct that it is supposed to measure than that the construct is correlated with other construct(s)	Peter (1981), McDonald (1999)
$T = \sqrt{AVE_{Vi}}$	Theoretical: target measure has a higher correlation with the construct that it is supposed to measure than the construct shares with other construct(s)	Fornell and Larcker (1981)

Notes: T refers to threshold, r_{ViVi} is the reliability of measure V_i, and AVE refers to the average variance extracted (Fornell and Larcker 1981).

The evidence for discriminant validity can then be dichotomous or continuous (Reichardt and Coleman 1995). The threshold is a sharp cutoff when heuristics are used. For instance, Farrell (2010, p. 325) interprets a squared correlation of .62 with T = .63 as evidence for discriminant validity, yet .67 relative to T = .65 does not establish discriminant validity, a dichotomous result. Treating the evidence for discriminant validity as continuous would judge the strength of the evidence based on the distance from the threshold. Using statistical tests, the dichotomous-continuous distinction might depend on the hypothesis testing paradigm (Perezgonzalez 2015). The Neyman-Pearson paradigm would either accept a hypothesis of discriminant invalidity or one of discriminant validity based on a *p*-value larger or smaller than a fixed value that is based on a-priori levels of confidence α and power 1-β. A Fisherian perspective would propose a null hypothesis of discriminant invalidity and calculate a *p*-value, which is then a continuous estimate of the strength of the evidence in the data, or evidential value, against the null hypothesis (Szucs and Ioannidis 2017a). Nevertheless, the common practice of null-hypothesis significance testing (NHST) that obfuscates principles of both Neyman-Pearson and Fisher paradigms (Hubbard 2019; Perezgonzalez 2015; Szucs and Ioannidis 2017a), is prone to meaningless null-hypotheses (Nickerson 2000) or thresholds. Importantly, misconceptions and misuse of *p*-values (Greenland et al. 2016) are likely to result in discrete discriminant validity testing. In practice, NHST might be suitable to make discrete decisions (Frick 1996), here: in case of discriminant validity, proceed with analysis; in case of discriminant invalidity, do not proceed with analysis. It is then essential to report correlation estimates as well as confidence intervals, absolute *p*-values, and to choose careful wording in interpretation to convince a reader of the statistical evidence for discriminant validity (Nickerson 2000).[7]

[7] *P*-values might also be transformed to power estimates (Hoenig and Heisey 2001) or Bayes-factors (Held and Ott 2018), yet those are also prone to ad hoc rules of thumb and cutoffs like *p*-values are, particularly when NHST-like principles are applied.

4.3.2 Multivariate discriminant validity (MDV).

The BDV criterion can be readily generalized to the multivariate case. Panel B in Figure 4.2 contains a graphical representation of a variance decomposition of V_1 with two other measures of constructs, V_2 and V_3. Segment a is the coefficient of alienation. The coefficient of determination is segments b + c + d. Segment b represents the *unique* variance contribution of V_2 to V_1. It is zero or positive as it is the squared semi-partial correlation of V_1 with V_2, partialing out V_3 (Cohen et al. 2003). Segment d is the squared semi-partial correlation of V_1 with V_3 (partialing out V_2). Segment c is the joint contribution of V_2 and V_3 to the variance of V_1. It cannot be interpreted as variance shared, as it can take negative values, for example in case of suppression (Cohen et al. 2003; Friedman and Wall 2005).

The multiple coefficient of determination is:

$$R^2_{Vi} = \mathbf{r}^{T}_{Vi,Vj}\mathbf{r}^{-1}_{Vj,Vj}\mathbf{r}_{Vi,Vj}, \tag{4.2}$$

where $\mathbf{r}_{ViVj}$ is a vector of correlations between the target measure of construct V_i and the other measures V_j, and $\mathbf{r}_{VjVj}$ is a matrix of correlations between the other measures V_j (superscript T and -1 denote respectively the matrix transpose and inverse). In case of three measures of constructs, the multiple coefficient of determination of V_1 with respect to V_2 and V_3 simplifies to:

$$R^2_{V1.V2,V3} = \frac{r^2_{V1,V2} + r^2_{V1,V3} - 2r_{V1,V2}r_{V1,V3}r_{V2,V3}}{1 - r^2_{V2,V3}}, \tag{4.3}$$

where $r_{Vi,Vj}$ refers to the correlation between V_i and V_j (with $i \neq j$). The notation in the sequel is $R_{V1.V2,V3}$ for the multiple correlation of V_1 with V_2-V_3, and $R_{V1.}$ for the multiple correlation of V_1 with all other measures. A criterion for MDV is then:

$$R_{Vi.} < T. \tag{4.4}$$

The criterion in Equation 4.4 captures whether the multiple correlation is smaller than a threshold. As in the bivariate case, correlation-based and reliability-based thresholds can be

used. For example, $R_{V1.}$ can be tested against a threshold of unity, T = 1, or the reliability of V_1, T = $r_{Vi,Vi}$. The remainder of this chapter explores both thresholds.

The key difference between the BDV and MDV criteria is that the BDV criterion relies on the correlation within a pair of measures, whereas the MDV criterion takes all correlations between the measures in a given set into account. The MDV criterion accounts for situations where pairwise associations fail to detect high shared variance of a focal measure with multiple other measures of constructs. For example, consider the case where a focal measure of a construct V_1 is correlated .70 with V_2 and .70 with V_3, and that V_2 and V_3 are uncorrelated. The two measures V_2 and V_3 then almost perfectly account for V_1 ($R^2_{V1.} \approx$ 98%), whereas the largest bivariate R^2 is 49%. Note that the multiple correlation is almost always higher than the bivariate correlations of the focal measure because its minimum is the highest correlation of the focal measure with the other measures.[8] For example, when $r_{V1,V2}$ = .80 and $r_{V1,V3}$ = .30, the multiple correlation takes its maximum value $R_{V1.}$ = 1 at $r_{V2,V3}$ = -.33, the minimum value. When $r_{V2,V3}$ increases, $R_{V1.}$ decreases until its minimum is reached, here at $r_{V1,V3}$ / $r_{V1,V2}$ = .30 / .80 = .375. $R_{V1.}$ then increases to 1 when $r_{V2,V3}$ further increases to about .81, the maximum.

A second strength of the MDV criterion is that it requires a smaller number of MDV tests than BDV tests when there are more than three constructs. If k is the number of constructs, the number of pairs for which BDV has to be established is k×(k-1) / 2, whereas there are only k MDV tests. The number of pairs increases exponentially with k which makes it inconvenient to establish BDV. Moreover, the higher likelihood of making false inferences increases when repeatedly using BDV due to the higher number of required statistical tests.

[8] The standard error of the multiple correlation is equal to that of the bivariate correlation: $\frac{1-R_{Vi.}}{\sqrt{df}}$, where df = n-k-1, and where n is the sample size and k is the number of measures in the set (Burt 1943; Isserlis 1917; Kelley 1932). The standard error is therefore equal for bivariate and multiple correlations except for the degrees of freedom. This difference is negligible in large samples.

For example, Fürst et al. (2017) analyzed survey-data of informants in 329 companies and established BDV between at least 10 constructs, which yielded 45 pairs of constructs to establish BDV. Establishing MDV would involve assessing 10 multiple correlations.

4.3.3 The impact of measurement error on BDV and MDV.

Thus far, we assumed that the measures of constructs do not contain measurement error. Random error in measures of constructs has a threefold impact on establishing multivariate discriminant validity. First, measurement error decreases the reliability of a measure and hence lowers the threshold of the reliability-based criterion if it is used. Second, measurement error increases the variance of parameters (Westfall and Yarkoni 2016), which results into wider confidence intervals that increase the likelihood that the confidence interval of the multiple correlation overlaps the threshold. Third, measurement error, when not accounted for, attenuates true bivariate correlations between measures of constructs toward zero which may lead to the erroneous conclusion that there is BDV whereas there is none. If the uncorrected correlation and reliabilities are known, the corrected (population) correlation can be readily established (Spearman 1904):

$$\hat{r}_{ViVj} = r_{ViVj} \times \sqrt{r_{ViVi} \times r_{VjVj}}, \qquad (4.5)$$

where $\hat{r}_{ViVj}$ is the uncorrected correlation and r_{ViVj} is the true correlation between constructs i and j, and r_{ViVi} and r_{VjVj} are the respective reliabilities of the measures for constructs i and j. Measurement error also attenuates the multiple correlation, albeit in a more complex fashion than for the bivariate case.

Table 4.2 illustrates this by presenting several scenarios of measurement error for a set of three measures of constructs. It has the corrected and uncorrected bivariate correlations and multiple correlation of V_1 for different values of $r_{V2,V3}$ (from 0 to .80), when $r_{V1,V2}$ = .80 and $r_{V1,V3}$ = .30. The reliabilities are fixed to either 1 (no measurement error) or .80 (measurement error is .20). The table also contains ratios of corrected / uncorrected

Table 4.2
Four Scenarios Illustrate that Measurement Error Attenuates the Bivariate and Multiple Correlations

				Bivariate correlations between V_1, V_2, and V_3									Multiple correlation of V_1		
	Reliability (1 – measurement error)			Uncorrected			Corrected			Corrected / uncorrected			Uncorrected	Corrected	Corrected / uncorrected
Scenario	$r_{V1,V1}$	$r_{V2,V2}$	$r_{V3,V3}$	$r_{V1,V2}$	$r_{V1,V3}$	$r_{V2,V3}$	$r_{V1,V2}$	$r_{V1,V3}$	$r_{V2,V3}$	$r_{V1,V2}$	$r_{V1,V3}$	$r_{V2,V3}$	$R_{V1.V2,V3}$	$R_{V1.V2,V3}$	$R_{V1.V2,V3}$
Scenario 1:	.80	1	1	.716	.268	0	.80	.30	0	1.118	1.118	1	.764	.854	1.118
Measurement						.20			.20				.727	.813	
error in						.375			.375				.716	.80	
V_1						.50			.50				.723	.808	
						.80			.80				.877	.980	
Scenario 2:	1	.80	1	.716	.30	0	.80	.30	0	1.118	1	1.118	.776	.854	1.101
Measurement						.179			.20				.737	.813	1.103
error in						.335			.375				.718	.80	1.114
V_2						.447			.50				.716	.808	1.129
						.716			.80				.777	.980	1.261
Scenario 3:	1	1	.80	.80	.268	0	.80	.30	0	1	1.118	1.118	.844	.854	1.013
Measurement						.179			.20				.810	.813	1.003
error in						.335			.375				.80	.80	1
V_3						.447			.50				.806	.808	1.003
						.716			.80				.911	.980	1.076
Scenario 4:	.80	.80	.80	.64	.24	0	.80	.30	0	1.250	1.250	1.250	.684	.854	1.250
Measurement						.16			.20				.655	.813	1.241
error in						.30			.375				.642	.80	1.246
V_1, V_2 & V_3						.40			.50				.640	.808	1.262
						.64			.80				.677	.980	1.448

correlations. For example, a ratio of 1.2 means that the corrected correlation is 20% larger than the uncorrected correlation. A ratio below one would mean that the corrected correlation is smaller than the uncorrected correlation.

Scenario 1 has measurement error in V_1. In this case, the ratio of corrected / uncorrected multiple correlations is about 1.118 = $1/\sqrt{.80}$, which is equal to the ratios for $r_{V1,V2}$ and $r_{V1,V3}$. Scenario 2 has measurement error in V_2. The ratio now varies, depending on $r_{V2,V3}$. For example, when $r_{V2,V3}$ is 0, there is about a 10% upward and when $r_{V2,V3}$ is .80, the upward correction is about 26%. Scenario 3 assumes measurement error in V_3 only. Across the range of $r_{V2,V3}$, the bias is less than in scenarios 1 and 2 because the correlation of V_1 with V_3 of .30 is relatively small compared to the correlation between V_1 with V_2 of .80. Scenario 4 considers measurement error in V_1, V_2 and V_3. Again, the bias varies over the range of $r_{V2,V3}$. When $r_{V2,V3} = 0$, the attenuation of course becomes equal to the bivariate case and the multiple correlation is only affected by the unreliability in V_1 and V_2. In conclusion, when measurement error is unaccounted for, there is a downward bias in the bivariate and multiple correlation, but the magnitude in the multiple correlation quickly becomes hard to establish analytically. It depends on the measures that are affected by measurement error (here reflected in the four scenarios) and the correlations between them.

In sum, when measurement error is unaccounted for, there is attenuation towards zero for both the bivariate and multiple correlations. Accounting for measurement error is paramount to make correct BDV and MDV inferences. Discriminant validity assessments without accounting for measurement error may lead to the erroneous conclusion that measures of constructs express discriminant validity although they truly do not. A Monte Carlo simulation study further explores and quantifies the consequences of measurement error, as well as other factors, on the MDV criteria.

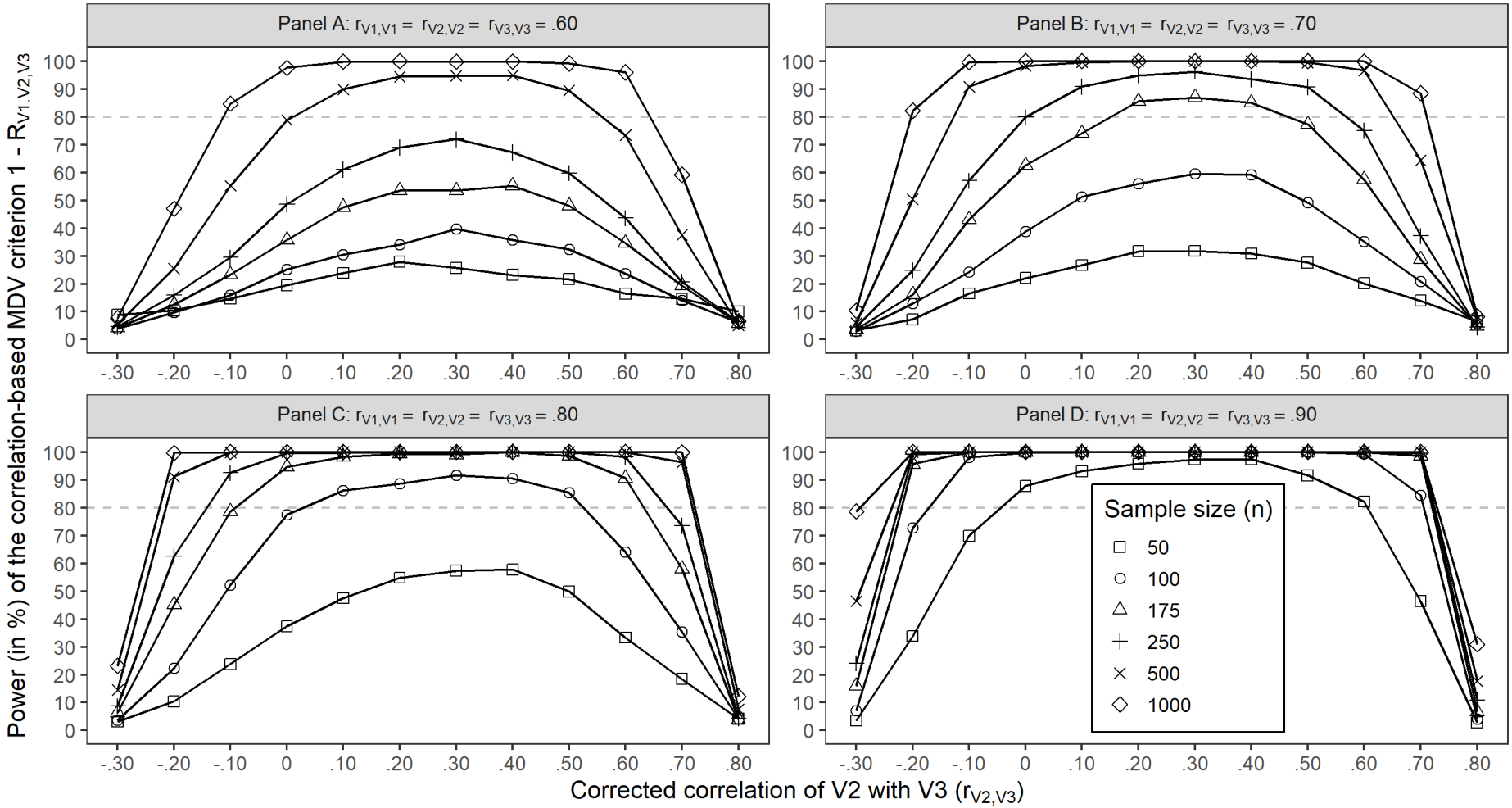
4.3.4 Power curves of MDV.

An estimated (corrected) multiple correlation of unity provides perfect evidence for the absence of MDV and a (corrected) correlation of zero indicates perfect evidence for the presence of MDV, regardless of the reliability and sample size. The key question is how unreliability (even when it is accounted for) and sample size influence the empirical evidence for MDV between these two boundaries. To investigate this, we conducted a Monte Carlo simulation study. The analysis had three measures of constructs with focal measure V_1. The correlations were fixed to $r_{V1,V2} = .80$, and $r_{V1,V3} = .30$ and $r_{V2,V3}$ varied from -.30 to .80. Reliability was fixed to .60 (poor), .70 (acceptable), .80 (good) and .90 (excellent) (Peterson 1994). The sample size varied from 50 (very small) to 1,000 (very large) (Pieters 2017). In sum, the full-factorial design consisted of 12 ($r_{V2,V3}$: -.30 to .80 in steps of .10) × 4 (reliability: .60, .70, .80 and .90) × 6 (sample size: 50, 100, 175, 250, 500, 1,000) = 288 cells.

We used structural equation modeling (SEM) implemented on the R platform for the analysis (R Core Team 2019; Rosseel 2012). The population model was a measurement model with three latent constructs. Factor loadings were fixed to one, and the error variances of the indicators were fixed to $3 \times (1-r_{Vi,Vi})/r_{Vi,Vi}$ to determine the reliability (Grewal et al. 2004). For each cell, we generated 1,000 replications and estimated the power of the correlation-based and reliability-based MDV criteria by calculating the proportion of replications for which the respective criterion was met.

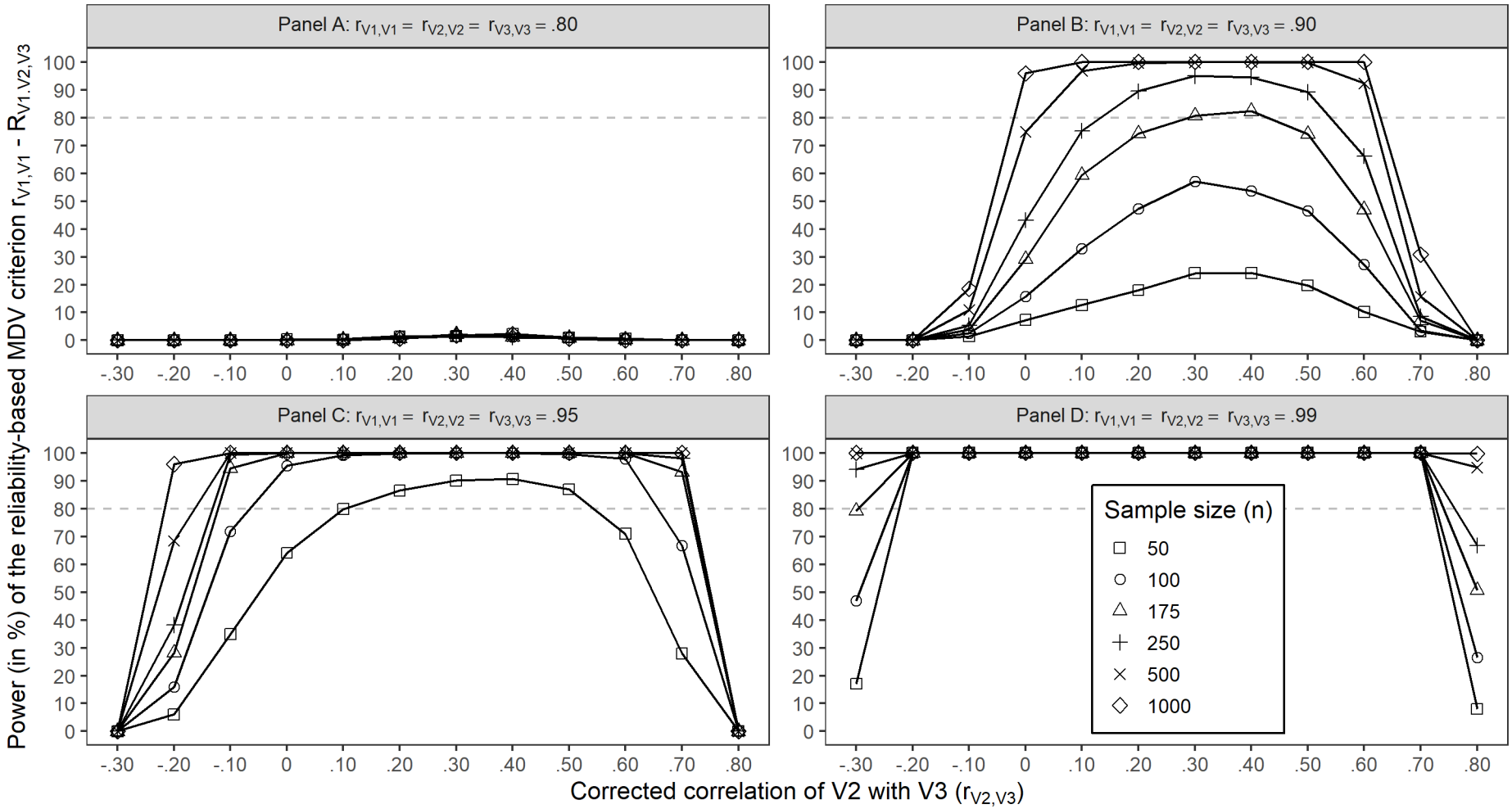
Figure 4.3 summarizes the estimated power curves for the correlation-based threshold ($R_{V1.V2,V3} < 1$) and Figure 4.4 has the power for the reliability-based criterion ($R_{V1.V2,V3} < R_{V1,V1}$). Both figures visualize the power of MDV for different reliabilities across the range of $r_{V2,V3}$. Figure 4.4 focuses on reliabilities .80, .90, .95 and .99 because reliabilities below .80 cannot meet the reliability-based criterion in these simulations because the minimum $R_{V1.}$ for the correlations in this design was .80. First, the results show an inverse U-shaped

Figure 4.3
Power Curves of the Correlation-Based Multivariate Discriminant Validity (MDV) Criterion of V_1



Notes: Results are based on a Monte Carlo simulation with 1,000 replications. Panels visualize the estimated power of the correlation-based MDV criterion (T = 1) of V_1 across $r_{V2,V3}$ (varied from -.30 to .80) for different reliabilities (.60 to .90), $r_{V1,V2}$ was fixed to .80 and $r_{V1,V3}$ was .30. The dashed line in each panel indicates 80% power.

Figure 4.4
Power Curves of the Reliability-Based Multivariate Discriminant Validity (MDV) Criterion of V_1



Notes: Results are based on a Monte Carlo simulation with 1,000 replications. Panels visualize the estimated power of the reliability-based MDV criterion ($T = r_{V1,V1}$) of V_1 across $r_{V2,V3}$ (varied from -.30 to .80) for different reliabilities (.80 to .99), $r_{V1,V2}$ was fixed to .80 and $r_{V1,V3}$ was .30. The dashed line in each panel indicates 80% power.

relationship between $r_{V2,V3}$ and the power of MDV. At both edges of the range of $r_{V2,V3}$, $R_{V1.}$ approaches one which leads to lower power of discriminant validity. The power is highest at the minimum of R_{V1}, here $r_{V2,V3} = .375$. Second, as expected, larger reliabilities increase the power of MDV.

A sample size of 175, which is about the median sample size in mediation analyses in recent consumer research (Pieters 2017), yields adequate power for the correlation-based MDV criterion for $r_{V2,V3} = .30$ and a reliability of .80 (estimated power of 100%), but has an estimated power of merely 50% if the reliability becomes .60 (Panels A and C of Figure 4.3). However, the influence of the reliability on statistical power is larger for the reliability-based criterion as it directly changes the threshold. Third, unsurprisingly, larger samples increase the power of MDV. In other words, larger samples are required for adequately powered MDV tests if reliability is low and the multiple correlation is high.

According to the simulation, small correlations of .10 or .20 can already lead to a power below 80% for a sample size of 175 and a high reliability of .90 (Panel B in Figure 4.4).

In sum, the Monte Carlo simulations show that high multiple correlations, low reliabilities (particularly for the reliability-based criterion), and small sample sizes decrease the power of discriminant validity. However, the question remains to what extent these conditions occur in real-world applications.

4.4 Discriminant Validity: The Case of Multiple Mediation

A literature review investigated discriminant validity in multiple mediation models. Such theories are common in consumer research (Deighton et al. 2010; Pieters 2017) and set out to identify and quantify the multiple pathways through which an input variable (X) has a causal effect on a relevant outcome (Y) through multiple mediators (M). They decompose the causal effect by which X leads to Y in multiple smaller steps (sequential or serial mediation),

separate chains (parallel mediation), or a combination of the two (sequential and parallel mediation). Empirical distinctiveness of mediators and outcomes is an essential condition for meaningful mediation analysis (Pieters 2017; Spencer et al. 2005). Yet, multiple mediation models are at a particular risk for lack of discriminant validity. Input variables and mediators, as well as sequential mediators with each other, are by definition hypothesized to be related. Mediators in parallel may also be highly correlated because they are both a function of X, and can fail to express discriminant validity if they do not capture distinct processes.[9]

4.4.1 Method.

We searched all articles published in three recent volumes of the *Journal of Marketing* (*JM;* volumes 81 to 83), *Journal of Marketing Research* (*JMR;* volumes 54 to 56), *Journal of Consumer Research* (*JCR;* volumes 44 to 46) and *Marketing Science* (*MktSc*; volumes 36 to 38) for relevant keywords. Specifically, the search term was (parallel mediat OR sequential mediat OR multiple mediat OR serial mediat). Studies were eligible if the mediation model had at least two continuous mediators, reliability information was available for the measures of at least two mediators or outcomes. Studies were included when correlations between at least three measures of constructs were reported or could be inferred from other information in the reports.

This resulted in 23 studies from 15 articles (4 articles in *JM*, 4 in *JMR*, 7 in *JCR*, none in *MktSc*). Out of the 23 studies, 13 hypothesized sequential mediation, 5 had parallel mediation, and 5 had a process model with sequential and parallel mediation. All except two studies (Auh et al. 2019; Fürst et al. 2017) used a manipulated X in the process model and Martin et al. (2017) administered a treatment but used difference scores between a measure

[9] Mediation models by design provide a special case where the Y-equation is affected by multicollinearity. If multiple mediators are in parallel, they jointly affect Y together with X, while X and M are also hypothesized to be correlated. If multiple mediators are sequential, the first (proximal) M and X jointly predict the second (distal) M, and both M's and X influence Y. In this case, Y is predicted by three variables that are predicted to be non-zero correlated with each other.

pre vs. post treatment as focal X. Three out of the 23 studies provided raw data (Goenka and Van Osselaer 2019; Paley et al. 2018; Steffel and Williams 2018). The remaining studies reported correlations or sufficient information to infer correlations between the measures. Correlations were inferred by transforming reported exact t-statistics of directed paths in the model to their underlying partial correlations and then to zero-order correlations, or by directly transforming overall F-statistics to correlations (Rosenthal and DiMatteo 2001).

Mean reported reliabilities and correlations were determined based on meta-analysis.[10] We transformed the estimates to Fisher-Z-values, took the mean of the Z-values, and back-transformed these to a meta-analytic mean correlation or reliability (Charter and Larsen 1983). Table 4.3 reports the simple and weighted means. The weight was the inverse of the standard error of the Z-values $\sqrt{n-2}$, where n is the sample size. It gives more weight to correlations from larger studies. The correlations were corrected for measurement error (Equation 4.5). However, reliability information was not always available (of 21 manipulated X, 2 single-indicator M and 12 single-indicator Y). In all cases, treatments were assumed to be without measurement error and reliabilities of one were imputed. Reliabilities for the missing single-indicators of M and Y were imputed by estimating their single-item reliabilities with the Spearman-Brown prophecy formula: $r_{Vi,Vi}^{single} = \frac{r_{Vi,Vi}}{r_{Vi,Vi} + u(1 - r_{Vi,Vi})}$, where $r_{Vi,Vi}^{single}$ is the estimated single-indicator reliability, $r_{Vi,Vi}$ the provided reliability of the multi-indicator scale, and u the number of indicators of the measure with multiple indicators, assuming equally good indicators (Pieters 2017). We used $r_{Vi,Vi}$ = .90 and u = 3 (about the median reliability and number of indicators for M and Y) which resulted in an estimated (high) single-indicator reliability of .75, which just exceeds the minimum threshold of .70

[10] Correlations between binary variables (here: treatments or manipulations) and continuous variables (here: mediators and outcomes) are point-biserial correlations which are equal to Pearson correlations that are used when both variables are continuous (Cohen et al. 2003).

recommended by Wanous and Hudy (2001). It is consistent with the observation that single-indicator measures often have lower reliabilities than scales with multiple indicators, although it is still comparatively high (Petrescu 2013; Wanous and Hudy 2001; Westfall and Yarkoni 2016).

4.4.2 Results.

As shown in Table 4.3, the median sample size across the studies was 200. The smallest sample size was 78. Studies had a median of one input (X), two mediators (M), and a single outcome (Y), with three studies having four mediators, and two studies having three outcomes. Measures of X had a median of one indicator, reflecting the large proportion of manipulations, and M and Y both had a median of three indicators. The reliabilities of X, M and Y were respectively .85, .91, and .90. The reliability of X was only reported for four measures and Y could only be assessed for 17 out of 27 measures. These reliabilities are slightly higher than the mean reliability of .84 for M and .85 for Y reported by Pieters (2017). Still, the lowest reliability was .68 for M and .77 for Y.

The uncorrected bivariate correlation between X and M was a median .26, and .20 between X and Y. M and Y were correlated higher: a median .57 for M with M, .48 for Y with Y, and .47 for M with Y. The corrected correlation between X and M was a median .28, and .22 between X and Y. This is in line with the mean effect size of .24 in a meta-analysis of meta-analyses in marketing (Eisend 2015). The median corrected correlations were .62 for M with M, .63 for Y with Y, and .51 for M with Y. It is expected that these correlations are higher than the correlations with X as X is often manipulated and correlated M and Y are selected in order to maximize the likelihood that mediation is supported.

The multiple correlations within a stage could not be assessed for X as no study had more than 2 X-measures. The median multiple correlation was a median of .71 for M and .69 for Y. Multiple correlations between stages were a median .38 for X, and higher for M and Y

Table 4.3
Parallel and Sequential Mediation Models Published in Three Volumes of *JM*, *JMR* and *JCR*

Category	Result
Sample size in mediation analysis (23 out of 23 studies)	Mean = 256, Mdn = 200 (SD = 209, range = 78-891)
Input (X)	24
Number per study (23 out of 23 studies)	Mean = 1.04, Mdn = 1 (SD = .21, range = 1-2)
Number of indicators (24 out of 24 X)	Mean = 1.58, Mdn = 1 (SD = 1.35, range = 1-5)
Reliability (4 out of 24 X)	Mean = .89, $Mean_w$ = .88, Mdn = .85 (SD = .07, range = .81-.97)
Mediators (M)	56
Number per study (23 out of 23 studies)	Mean = 2.44, Mdn = 2 (SD = .73, range = 2-4)
Number of indicators (56 out of 56 M)	Mean = 3.93, Mdn = 3 (SD = 2.26, range = 1-13)
Reliability (54 out of 56 M)	Mean = .91, $Mean_w$ = .91, Mdn = .91 (SD = .07, range = .68-.98)
Outcomes (Y)	27
Number per study (23 out of 23 studies)	Mean = 1.17, Mdn = 1 (SD = .58, range = 1-3)
Number of indicators (27 out of 27 Y)	Mean = 2.70, Mdn = 3 (SD = 1.75, range = 1-7)
Reliability (17 out of 27 Y)	Mean = .91, $Mean_w$ = .91, Mdn = .90 (SD = .05, range = .77- 97)
Bivariate correlations - uncorrected	
Within stage	
X with X (1 out of 1 correlation)	.20
M with M (46 out of 46 correlations)	Mean = .55, $Mean_w$ = .53, Mdn = .57 (SD = .16, range = .18-.78)
Y with Y (6 out of 6 correlations)	Mean = .49, $Mean_w$ = .50, Mdn = .48 (SD = .09, range = .33-.60)
Between stage	
X with M (53 out of 59 correlations)	Mean = .30, $Mean_w$ = .29, Mdn = .26 (SD = .16, range = .01-.65)
X with Y (26 out of 28 correlations)	Mean = .23, $Mean_w$ = .21, Mdn = .20 (SD = .12, range = .08-.54)
M with Y (56 out of 64 correlations)	Mean = .48, $Mean_w$ = .42, Mdn = .47 (SD = .23, range = .03-.88)
Bivariate correlations - corrected	
Within stage	
X with X (1 out of 1 correlation)	.25
M with M (46 out of 46 correlations)	Mean = .60, $Mean_w$ = .59, Mdn = .62 (SD = .15, range = .23-.79)
Y with Y (6 out of 6 correlations)	Mean = .59, $Mean_w$ = .63, Mdn = .63 (SD = .14, range = .35-.74)
Between stage	
X with M (53 out of 59 correlations)	Mean = .32, $Mean_w$ = .31, Mdn = .28 (SD = .16, range = .01-.69)
X with Y (26 out of 28 correlations)	Mean = .25, $Mean_w$ = .23, Mdn = .22 (SD = .12, range = .10-.54)
M with Y (56 out of 64 correlations)	Mean = .51, $Mean_w$ = .46, Mdn = .51 (SD = .23, range = .04-.91)
Multiple correlations - corrected	
Within stage	
X with X ($R_{X.X}$; 0 out of 24 X)	-
M with M ($R_{M.M}$; 24 out of 56 M)	Mean = .71, $Mean_w$ = .71, Mdn = .71 (SD = .10, range = .44-.84)
Y with Y ($R_{Y.Y}$; 6 out of 27 Y)	Mean = .68, $Mean_w$ = .71, Mdn = .69 (SD = .11, range = .48-.78)
Between stage	
X with X & M & Y ($R_{X.X,M,Y}$; 20 out of 24 X)	Mean = .46, $Mean_w$ = .47, Mdn = .38 (SD = .16, range = .19-.71)
M with X & M & Y ($R_{M.X,M,Y}$; 56 out of 56 M)	Mean = .74, $Mean_w$ = .74, Mdn = .74 (SD = .13, range = .30-.92)
Y with X & M & Y ($R_{Y.X,M,Y}$; 25 out of 27 Y)	Mean = .70, $Mean_w$ = .70, Mdn = .68 (SD = .14, range = .40-.92)

Notes: Mean is the arithmetic (simple) mean and $Mean_w$ is the weighted (by inverse of the standard error) mean, Mdn is the median, and SD refers to the standard deviation. For reliabilities and correlations, the reported (weighted) means are back-transformed (weighted) means of Fisher-Z-transformed values. Reported means of bivariate correlations are computed using absolute values. Missing reliabilities (to compute corrected correlations) were imputed 1 for manipulated X and .75 for single-indicator M and Y. The numbers in parentheses in the first column denote the number of studies, constructs, or correlations that the corresponding statistics in the second column are based on because some data could not be unequivocally determined from the study reports.

(respectively .74 for M, and .68 for Y). The maximum multiple correlation is our sample is a very high .92. These results suggest that at the median about half of the variance in mediators and outcomes is accounted for by the other mediators and outcomes ($.70^2 \times 100\% = 49\%$). Moreover, because M and Ys have higher multivariate correlations than X, they are at a higher risk of failing to express discriminant validity than X is.

Despite a high median multiple correlation, discriminant validity is not reported on in the majority of the 23 studies, a conclusion that is similar to what has been found elsewhere (Pieters 2017; Voorhees et al. 2016). Table 4.4 contains data on the reported discriminant validity in the 15 articles that we examined. Five report BDV and, unsurprisingly, none MDV. Some of the articles use statistical criteria, such as chi-square difference tests (Fürst et al. 2017), to assess discriminant validity. But none report statistical evidence to support the claim of discriminant validity. An exception is the reporting of confidence intervals not overlapping one in support of the correlation-based criterion in Bellezza et al. (2017). The others rely on heuristics in support of discriminant validity (Franke and Sarstedt 2018). The remaining 10 articles did not report on discriminant validity.

In sum, discriminant validity is rarely reported. Yet, we find evidence for high multiple correlations between measures of constructs. In what follows, four case studies illustrate real-world conditions in which discriminant validity is unlikely to be established.

4.4.3 Case studies.

We investigate four cases. First, we assess the discriminant validity at the median values from the meta-analyses. Then, three reanalyses zoom in on discriminant validity of three of the 23 individual studies. For illustration, these focus on cases that are at the highest risk of not establishing discriminant validity. We selected the studies with the highest multiple correlation ($R = .92$, Case 2), the lowest reliability ($r_{Vi,Vi} = .68$, Case 3), and the smallest sample size ($n = 78$; Case 4). We used structural equation modeling and Monte Carlo

simulation methods with 1,000 replications (Muthén and Muthén 2002). For each replication, the BDV and MDV criteria were met if the difference between the correlation and threshold was larger than zero. The proportion of statistically significant differences that was positive and larger than zero is an estimate of statistical power. We studied correlation-based and reliability-based criteria. We consider an estimated power for the correlation-based (T = 1) and reliability-based criteria higher than 80% to be evidence for discriminant validity (Cohen 1988). If the power of discriminant validity is low, it is unlikely that discriminant validity, if it was found, replicates.

Case 1: Median values from the meta-analysis.

Panel A in Table 4.5 shows the summary statistics data (SSD) based on the median values from the meta-analysis. The correlations between measures were small, the highest was .57 between M_1 and M_2. Reliabilities were high (above .90). At the median sample size, n = 200, all measures express BDV (Panels B and C) and MDV (Panel D) based on the correlation-based and reliability-based criteria (estimated power = 100%). Thus, studies at the median of

Table 4.4
Reported Discriminant Validity in 15 Articles with Multiple Mediation Analyses in Marketing Research

(A) Article	(B) DV	(C) BDV / MDV	(D) C-B	(E) R-B	(F) Measurement error accounted for	(G) Statistical evidence reported
Auh et al. (2019)	No	-	-	-	-	-
Bellezza et al. (2017)	Yes	BDV	Yes (T = 1)	Yes (AVE>VS)	Yes	Only for C-B
Eggert et al. (2019)	Yes	BDV	No	Yes (AVE>VS)	Yes	No
Fürst et al. (2017)	Yes	BDV	Yes (T = 1)	Yes (AVE>VS)	No	No
Goenka and Van Osselaer (2019)	No	-	-	-	-	-
Grewal and Stephen (2019)	No	-	-	-	-	-
Huyghe et al. (2017)	No	-	-	-	-	-
Martin et al. (2017)	No	-	-	-	-	-
Müller-Stewens et al. (2017)	Yes	BDV	No	Yes (AVE>VS)	Yes	No
Paley et al. (2018)	No	-	-	-	-	-
Schroll et al. (2018)	Yes	BDV	No	Yes (AVE > .50 & AVE>VS)	Yes	No
Shen and Sengupta (2018)	No	-	-	-	-	-
Steffel and Williams (2018)	No	-	-	-	-	-
Van Laer et al. (2018)	No	-	-	-	-	-
Wang et al. (2017)	No	-	-	-	-	-

Notes: Column B refers to whether discriminant validity was reported on. In Column C, BDV refers to bivariate discriminant validity, MDV is multivariate discriminant validity. Column D contains whether the correlation-based (C-B) was reported, with the threshold(s) used in parentheses. Column E reports whether a reliability-based criterion was reported, with the threshold within parentheses. AVE>VS refers to a comparison of the average variance extracted with the variance shared (Fornell and Larcker 1981). In Column E, statistical evidence means that statistical tests are reported (e.g., confidence intervals that do not overlap one or difference tests between correlations and reliabilities) instead of heuristics.

the meta-analysis have a relatively low risk to fail the BDV and MDV criteria. The remaining cases zoom in on situations that are less likely to express discriminant validity.

Case 2: High multiple R – Study 3 in Eggert et al. (2019).

Study 3 in Eggert et al. (2019) hypothesized a process model with both parallel and serial mediation. Receiving assistance (X_1) and branded gift wrapping (X_2) strengthen purchase intention (Y) after purchasing gifts via gratitude (M_1) and public commitment (M_2), which both lead to increased attitude strength (M_3). An online experiment (n = 159) let participants envision a scenario with manipulations for X_1 and X_2.

Panel A in Table 4.6 contains summary statistics data (SSD). The reanalysis focuses on the mediators M_1-M_3 and the outcome Y only because correlations with X could not be unequivocally determined from the report. The observed (uncorrected) correlations between mediators and the outcome were high, the highest is .87 between M_3 and Y. Nevertheless, the reliabilities were also high (minimum reliability was .94 for M_2). Panel B and Panel C show the results for BDV. The correlation-based criterion in this case is met, with estimated power of 100% for all correlations. The corrected correlation between M_3 and Y is a high .91 (95% CI [.88, .95]) and did not meet the reliability-based criterion (59% power for the difference with the reliability of M_2 and 88% for the difference with the reliability of Y). Similarly, as shown in Panel D of Table 4.6, the reliability-based MDV criterion was not met for M_3. The estimated multiple R was .92 (95% Monte Carlo CI [.89, .95]) for M_3 and Y, and the estimated power for M_3 was 32%, and 80% for Y. This difference in power is explained by the slightly higher reliability of Y (.96 for Y and .95 for M_2). Thus, Case 2 provides an example where MDV is not established due to high multiple correlations (here: .92). In sum, these results suggest that the measures for M_3 (attitude strength) and Y (purchase intention) may best be considered as a single measure for an underlying (positive) attitudes and intentions construct.

Case 3: Low reliability – Study 4 in Goenka and Van Osselaer (2019).

Goenka and Van Osselaer (2019) examine how evoking gratitude and compassion emotions leads to prosocial behavior (such as donating to a charity). Study 4 (n = 200 MTurk participants) tests the hypothesis that gratitude vs. compassion (X, manipulated) influences moral concerns, namely care (M_1) and fairness (M_2), which in parallel lead to a preference for a charity concerned with care in society or fairness in society (Y).

Panel A of Table 4.7 has SSD. Reliability information for the single-indicator Y was not available and its reliability was assumed to be .75, as indicated before. In this study, the largest uncorrected correlation was .29 (between M_1 and M_2). However, reliabilities were also low to moderate (minimum reliability of .68 for the measure of M_2). As shown in Panels B and C, all pairs meet the correlation-based and reliability-based BDV criteria (lowest estimated power 85% for corrected correlation M_1 with M_2, estimated .43 with 95% CI [.24, .57]). Panel D shows that the correlation-based MDV criterion was also met. However, M_2 failed to meet the reliability-based MDV criterion (estimated power = 20%). The multiple R for M_2 was an estimated .62 (95% Monte Carlo CI [.47, .77]) which is relatively low, but it did not significantly differ from its reliability (.68, 95% CI [.61, .75]). The results of this case suggest that although the reported mediation analysis supported the predictions, the support for the purported process model is weakened due to lack of MDV. It is an example where discriminant validity is not met due to low reliabilities of the measures, despite the reasonable sample size (median from the literature review) and low uncorrected correlations (highest uncorrected correlation was .29 which is much smaller than the median from the literature review).

Case 4: Small sample size – Study 5 in Shen and Sengupta (2018).

Shen and Sengupta (2018) study the impact of the communication channel of reviews on self-brand connection. The model specified that talking (X, as compared to writing) heightens the

extent to which communicators focus on the interaction aspect of the communication, which leads to increased interaction focus (M_1), and in sequence leads to higher self-expression (M_2) and self-brand connection (Y). The study had a 2 (communication channel) × 2 (prior interaction with the recipient of the communication) design. The mediation analysis only used data from the control condition of the prior interaction factor, which yielded a sample size for mediation analysis of 78 out of 153 participants (see Table 4, p. 502 in Shen and Sengupta 2018), which is less than half the median sample size of 200 in the literature review.

Panel A in Table 4.8 has the SSD. Although the correlation-based BDV criterion was met, the reliability-based criterion was not met for M_2 with Y: the estimated power was a mere 46% (estimated correlation was .62 with 95% CI [.44, .80]). M_2 did not meet the reliability-based MDV criterion (power = 20%) despite the moderate multiple R of .69. At the small sample size of 78, relatively small bivariate and multivariate correlations can fail to establish discriminant validity due to wide confidence intervals, as is the case here.

4.5 Online Implementation

Three cases demonstrated weak evidence for discriminant validity. Nevertheless, none of these cases reported on MDV, and only one on BDV (Table 4.4). A possible explanation for this result, and the overall scarce reporting on discriminant validity documented elsewhere (Pieters 2017; Voorhees et al. 2016), is the accessibility of discriminant validity analyses in conventional statistical software. We therefore developed an accessible online Shiny application to assess discriminant validity (Chang et al. 2019; R Core Team 2019). The application can be accessed at https://github.com/constantpieters/dv. The application estimates the power of discriminant validity at a given sample size (set n, estimate power) or the required sample size for a predetermined power level (set power, estimate n).

Table 4.5
Case 1: Meta-Analytic Median Values

Panel A: Summary Statistics Data (SSD)

		X Input	M_1 Mediator	M_2 Mediator	Y Outcome
X	Input	(1 / 1)			
M_1	Mediator	.26	(3 / .91)		
M_2	Mediator	.26	.57	(3 / .91)	
Y	Outcome	.20	.47	.47	(3 / .90)

Panel B: Bivariate Discriminant Validity (BDV) - Correlation-Based Criterion

		X Input	M_1 Mediator	M_2 Mediator	Y Outcome
X	Input	1	100%	100%	100%
M_1	Mediator	.27 [.14, .41]	1	100%	100%
M_2	Mediator	.27 [.14, .41]	.62 [.53, .72]	1	100%
Y	Outcome	.21 [.07, .35]	.52 [.40, .63]	.52 [.40, .64]	1

Panel C: Bivariate Discriminant Validity (BDV) - Reliability-Based Criterion

		X Input	M_1 Mediator	M_2 Mediator	Y Outcome
X	Input	(1)	100%; 100%	100%; 100%	100%; 100%
M_1	Mediator	.27 [.14, .41]	(.91) [.89, .93]	100%; 100%	100%; 100%
M_2	Mediator	.27 [.14, .41]	.63 [.53, .72]	(.91) [.89, .93]	100%; 100%
Y	Outcome	.21 [.07 .35]	.52 [.40, .63]	.52 [.40, .64]	(.90) [.88, .92]

Panel D: Multivariate Discriminant Validity (MDV) – Correlation- & Reliability-Based Criteria

		Estimated multiple R	Estimated reliability ($r_{Vi,Vi}$)	% Power correlation-based criterion (1 – R)	% Power reliability-based criterion ($r_{Vi,Vi}$ - R)
X	Input	.32 [.20, .45]	1	100%	100%
M_1	Mediator	.68 [.59, .76]	.91 [.89, .93]	100%	100%
M_2	Mediator	.68 [.59, .76]	.91 [.89, .93]	100%	100%
Y	Outcome	.59 [.48, .69]	.90 [.88, .92]	100%	100%

Notes: Median n = 200. Estimates are based on Monte Carlo simulations with 1,000 replications. Mean 95% confidence intervals are between brackets. Panel A has bivariate uncorrected correlations between the measures of the constructs and # indicators / reliability in parentheses on the diagonal, these SSD are based on median values from the literature review (Table 4.3). Panel B has estimated bivariate corrected correlations below the diagonal and power of the correlation-based criterion with T = 1 above the diagonal. Panel C has reliabilities on the diagonal in parentheses and power of the reliability-based criterion (T = $r_{Vi,Vi}$) with respect to the measures in the row and column (separated by a ;) above the diagonal.

Table 4.6
Case 2: High Multiple R – Parallel and Sequential Mediation Study 3 in Eggert et al. (2019)

Panel A: Summary Statistics Data (SSD)

		M_1 Customer gratitude	M_2 Public commitment	M_3 Attitude strength	Y Purchase intention
M_1	Customer gratitude	(4 / .95)			
M_2	Public commitment	.53	(3 / .94)		
M_3	Attitude strength	.75	.55	(1 / .95)	
Y	Purchase intention	.74	.58	.87	(4 / .96)

Panel B: Bivariate Discriminant Validity (BDV) - Correlation-Based Criterion

		M_1 Customer gratitude	M_2 Public commitment	M_3 Attitude strength	Y Purchase intention
M_1	Customer gratitude	1	100%	100%	100%
M_2	Public commitment	.56 [.44, .67]	1	100%	100%
M_3	Attitude strength	.79 [.73, .86]	.53 [.47, .69]	1	100%
Y	Purchase intention	.77 [.70, .84]	.61 [.50, .71]	.91 [.88, .95]	1

Panel C: Bivariate Discriminant Validity (BDV) - Reliability-Based Criterion

		M_1 Customer gratitude	M_2 Public commitment	M_3 Attitude strength	Y Purchase intention
M_1	Customer gratitude	(.95) [.94, .96]	100%; 100%	100%; 100%	100%; 100%
M_2	Public commitment	.56 [.44, .67]	(.94) [.92, .96]	100%; 100%	100%; 100%
M_3	Attitude strength	.79 [.73, .68]	.58 [.47, .69]	(.95)	59%; 88%
Y	Purchase intention	.76 [.70, .84]	.61 [.50, .71]	.91 [.88, .95]	(.960) [.95, .97]

Panel D: Multivariate Discriminant Validity (MDV) – Correlation- & Reliability-Based Criteria

		Estimated multiple R	Estimated reliability ($r_{Vi,Vi}$)	% Power correlation-based criterion (1 – R)	% Power reliability-based criterion ($r_{Vi,Vi}$ - R)
M_1	Customer gratitude	.81 [.75, .87]	.95 [.94, .96]	100%	100%
M_2	Public commitment	.63 [.53, .73]	.94 [.92, .96]	100%	100%
M_3	Attitude strength	.92 [.89, .95]	.95	100%	32%
Y	Purchase intention	.92 [.89, .95]	.96 [.95, .97]	100%	80%

Notes: n = 159. Estimates are based on Monte Carlo simulations with 1,000 replications. Mean 95% confidence intervals are between brackets. Panel A has bivariate uncorrected correlations between the measures of the constructs and # indicators / reliability in parentheses on the diagonal. We inferred the reliability of the product term M_3 as a function of the reliabilities of the components (Busemeyer and Jones 1983) assuming a correlation between the components of .71, which is the average of three correlations of valence with certainty in Park et al. (2010) from which Eggert et al. (2019) adapted the measures. Panel B has estimated bivariate corrected correlations below the diagonal and power of the correlation-based criterion with T = 1 above the diagonal. Panel C has reliabilities on the diagonal in parentheses and power of the reliability-based criterion ($T = r_{Vi,Vi}$) with respect to the measures in the row and column (separated by a ;) above the diagonal.

Table 4.7
Case 3: Low Reliability – Parallel Mediation Study 4 in Goenka and Van Osselaer (2019)

Panel A: Summary Statistics Data (SSD)

		X Emotion condition	M_1 Care	M_2 Fairness	Y Charity preference
X	Emotion condition	(1 / 1)			
M_1	Care	-.20	(6 / .78)		
M_2	Fairness	.15	.29	(6 / .68)	
Y	Charity preference	.16	-.19	.20	(1 / .75)

Panel B: Bivariate Discriminant Validity (BDV) - Correlation-Based Criterion

		X Emotion condition	M_1 Care	M_2 Fairness	Y Charity preference
X	Emotion condition	1	100%	100%	100%
M_1	Care	-.23 [-.37, -.08]	1	100%	100%
M_2	Fairness	.18 [.02, .34]	.40 [.24, .57]	1	100%
Y	Charity preference	.19 [.03, .40]	-.25 [-.42, -.08]	.28 [.01, .46]	1

Panel C: Bivariate Discriminant Validity (BDV) - Reliability-Based Criterion

		X Emotion condition	M_1 Care	M_2 Fairness	Y Charity preference
X	Emotion condition	1	100%; 100%	100%; 100%	100%; 100%
M_1	Care	-.23 [-.37, -.08]	(.78) [.73, .83]	100%; 85%	100%; 100%
M_2	Fairness	.18 [.02, .34]	.40 [.24, .57]	(.68) [.61, .75]	99%; 100%
Y	Charity preference	.19 [.03, .40]	-.25 [-.42, -.08]	.28 [.01, .46]	(.75)

Panel D: Multivariate Discriminant Validity (MDV) – Correlation- & Reliability-Based Criteria

		Estimated multiple R	Estimated reliability ($r_{Vi,Vi}$)	% Power correlation-based criterion (1 – R)	% Power reliability-based criterion ($r_{Vi,Vi}$ - R)
X	Emotion condition	.39 [.24, .55]	1	100%	100%
M_1	Care	.62 [.48, .76]	.78 [.73, .83]	100%	60%
M_2	Fairness	.62 [.47, .77]	.68 [.61, .75]	100%	11%
Y	Charity preference	.50 [.33, .67]	.75	100%	82%

Notes: n = 200. Estimates are based on Monte Carlo simulations with 1,000 replications. Mean 95% confidence intervals are between brackets. Panel A has bivariate uncorrected correlations between the measures of the constructs and # indicators / reliability in parentheses on the diagonal, based on raw data from https://osf.io/sczv4/. The reliability of .75 for the single-indicator Y was imputed with the Spearman-Brown formula. Panel B has estimated bivariate corrected correlations below the diagonal and power of the correlation-based criterion with T = 1 above the diagonal. Panel C has reliabilities on the diagonal in parentheses and power of the reliability-based criterion (T = $r_{Vi,Vi}$) with respect to the measures in the row and column (separated by a ;) above the diagonal.

Table 4.8
Case 4: Small Sample Size – Sequential Mediation Study 5 in Shen and Sengupta (2018)

Panel A: Summary Statistics Data (SSD)

		X Communication channel	M_1 Interaction focus	M_2 Self-expression	Y Self-brand connection
X	Communication channel	(1 / 1)			
M_1	Interaction focus	.26	(3 / .89)		
M_2	Self-expression	.22	.39	(1 / .80)	
Y	Self-brand connection	.21	.34	.52	(7 / .90)

Panel B: Bivariate Discriminant Validity (BDV) - Correlation-Based Criterion

		X Communication channel	M_1 Interaction focus	M_2 Self-expression	Y Self-brand connection
X	Communication channel	1	100%	100%	100%
M_1	Interaction focus	.27 [.05, .48]	1	100%	100%
M_2	Self-expression	.25 [-.02, .49]	.46 [.25, .68]	1	100%
Y	Self-brand connection	.22 [0, .44]	.37 [.16, .59]	.62 [.44,.80]	1

Panel C: Bivariate Discriminant Validity (BDV) - Reliability-Based Criterion

		X Communication channel	M_1 Interaction focus	M_2 Self-expression	Y Self-brand connection
X	Communication channel	(1)	100%; 100%	100%; 100%	100%; 100%
M_1	Interaction focus	.27 [.05, .48]	(.89) [.85, .93]	100%; 91%	100%; 100%
M_2	Self-expression	.25 [-.02, .49]	.46 [.25, .68]	(.80)	46%; 94%
Y	Self-brand connection	.22 [0, .44]	.37 [.16, .59]	.62 [.44, .80]	(.90) [.86, .93]

Panel D: Multivariate Discriminant Validity (MDV) – Correlation- & Reliability-Based Criteria

		Estimated multiple R	Estimated reliability ($r_{Vi,Vi}$)	% Power correlation-based criterion (1 – R)	% Power reliability-based criterion ($r_{Vi,Vi}$ - R)
X	Communication channel	.36 [.15, .57]	1	100%	100%
M_1	Interaction focus	.53 [.34, .72]	.89 [.85, .93]	100%	99%
M_2	Self-expression	.69 [.53, .85]	.80	100%	20%
Y	Self-brand connection	.65 [.48, .82]	.90 [.86, .93]	100%	92%

Notes: n = 78. Estimates are based on Monte Carlo simulations with 1,000 replications. Mean 95% confidence intervals are between brackets. Panel A has bivariate uncorrected correlations between the measures of the constructs and # indicators / reliability in parentheses on the diagonal. Reliabilities for M_1 and Y_1 were imputed with reliabilities from Study 1 and 3 in Shen and Sengupta (2018). The reliability estimate of M_2 is a reported interrater agreement. Panel B has estimated corrected bivariate correlations below the diagonal and power of the correlation-based criterion with T = 1 above the diagonal. Panel C has reliabilities on the diagonal in parentheses and power of the reliability-based criterion (T = $r_{Vi,Vi}$) with respect to the measures in the row and column (separated by a ;) above the diagonal.

The current version of the application supports up to four measures of constructs and reports BDV and MDV analyses. It uses SSD as input and estimates the power using Monte Carlo simulations of a single-indicator structural equation model (Bollen 1989; Jöreskog and Sörbom 1989; Muthén and Muthén 2002).

The application estimates the power of discriminant validity at a given sample size (set n, estimate power) or the required sample size for a predetermined power level (set power, estimate n). The current version of the application supports up to four measures of constructs and reports BDV and MDV analyses.[11] It uses SSD as input and estimates the power using Monte Carlo simulations of a single-indicator structural equation model (Bollen 1989; Jöreskog and Sörbom 1989; Muthén and Muthén 2002). An advantage of our approach is that it facilitates the analyses when raw data are not available, for example in study planning, meta-analysis, or evaluation of manuscripts in the review process.

Figure 4.5 has screen captures of the application. The Appendix of Chapter 4 contains additional details. It contains one of the available cases that can be readily used: discriminant validity of the different facets of market orientation (Narver and Slater 1990). It has three dimensions: customer orientation (understanding of one's target buyers), competitor orientation (understanding of the competitors) and interfunctional coordination (coordinated utilization of company resources in creating value for customers). Narver and Slater (1990) assessed discriminant validity by comparing the correlations between the three facets with the correlation of each facet with a fourth measure, of human resource management policy. The correlation between the facets should be higher than the correlation of each facet with human resource management policy, which is essentially a nomological discriminant validity criterion. Correlation- and reliability-based criteria were not reported on.

[11] The current version (v1.0.0, as of February 2020) implements the correlation-based criterion with a threshold of one. Future updates plan to include other criteria.

Panel A in Figure 4.5 shows the results of the reanalysis. The application has two sections. The left section has input and settings for the analysis and contains SSD reported in Tables 1 and 3 in Narver and Slater (1990, pp. 24-26) for measures of V_1 (customer orientation) V_2 (competitor orientation) and V_3 (interfunctional coordination). The sample size was n = 365, the uncorrected correlations between the facets were fairly high (highest .74 for V_1 with V_2) and the reliabilities were acceptable (minimum reliability .71 for V_3). The right section in Panel A of Figure 4.5 shows the output for BDV based on 1,000 Monte Carlo replications (the default). It shows a matrix with estimated corrected correlations below the diagonal, and statistical power of the correlation-based criterion with threshold T = 1 above the diagonal. The results show weak evidence for bivariate discriminant validity. The power of the correlation-based BDV criterion was 63% for V_1 with V_2, and 61% for V_2 with V_3, well below 80%. The estimated power was 77% for V_1 with V_3. Panel B in Figure 4.5 shows the results of the correlation-based MDV criterion. The second column shows the estimated corrected multiple correlations, and the third column has estimated power levels for the correlation-based MDV criterion. The estimated power was 58%, 48% and 56% for V_1-V_3, again well below 80%.

Located in the top right of Panel B in Figure 4.5, the app includes functionality to download a report of these results. In sum, there is weak support for discriminant validity between the three facets of market orientation in (Narver and Slater 1990). The results suggest that the three measures may best be aggregated to measures of overall market orientation.

This reanalysis used default settings, and the left section in Panel B in Figure 4.5 shows additional settings. The first setting changes the analysis from power estimation (set n, estimate power; the default) to sample size estimation (set power, estimate n). The sample size estimation routine uses a regression to estimate the required sample size for adequate

Figure 4.5

Online Shiny Application to Assess Discriminant Validity

Panel A: Input and Bivariate Discriminant Validity (BDV)

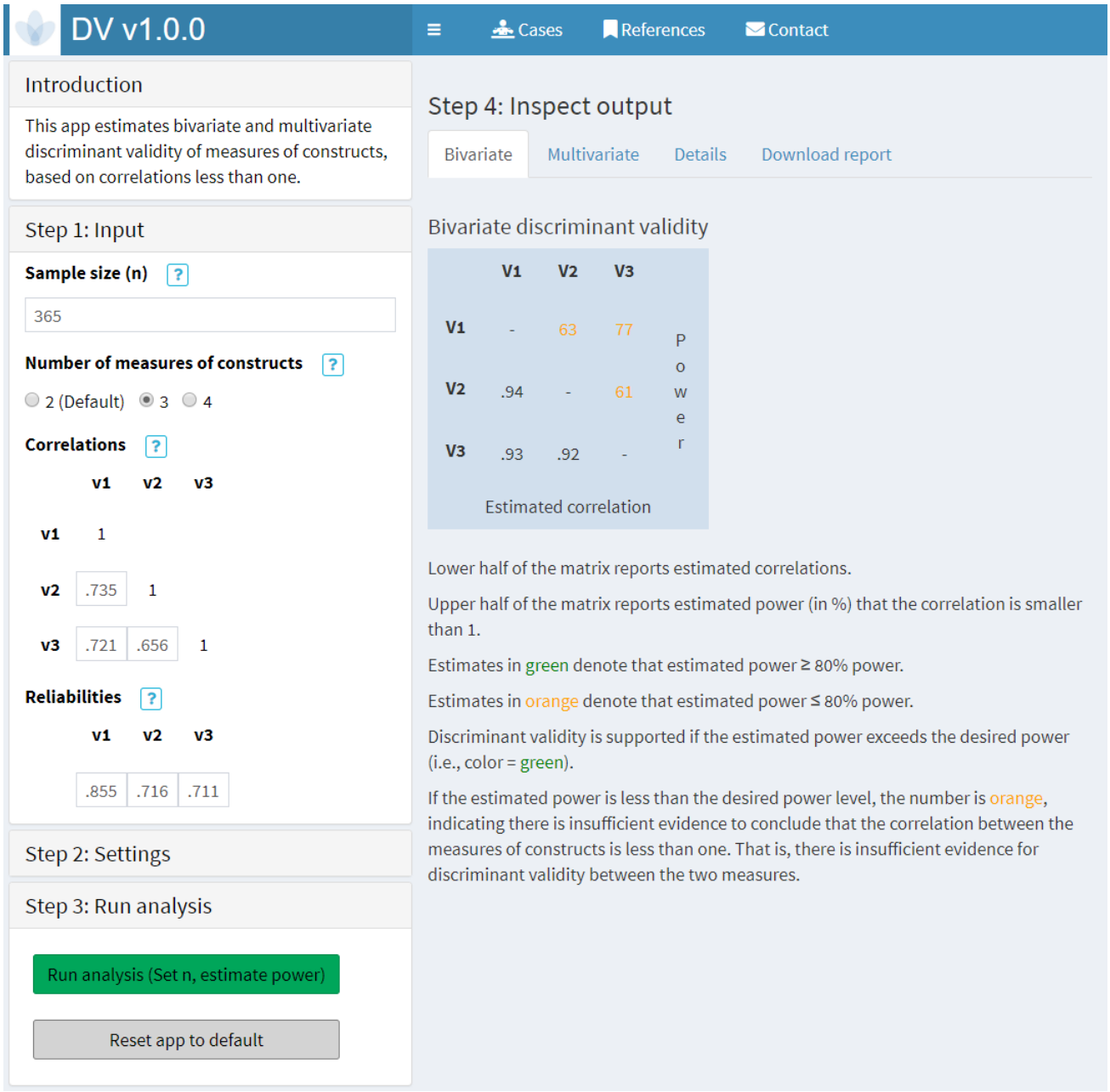
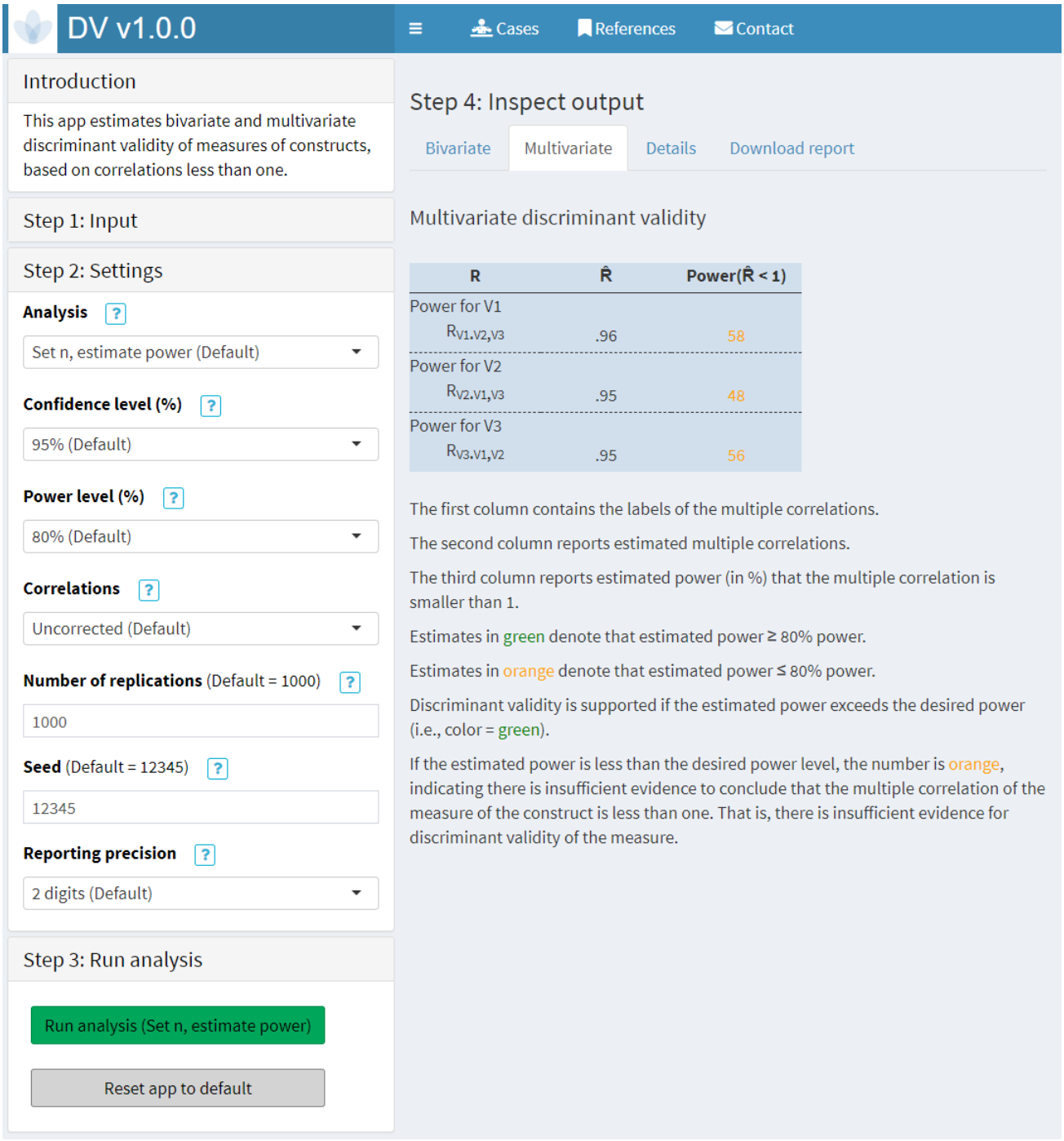


Figure 4.5 (CONTINUED)

Panel B: Settings and Multivariate Discriminant Validity (MDV)



Note: Screen captures contain the input and output for reanalysis of Narver and Slater (1990) as one of the readily available cases in the application at https://github.com/constantpieters/dv.

power based on a Monte Carlo simulation with varying sample sizes, as outlined in Schoemann et al. (2014). Additional settings modify the default confidence (90%, 95% or 99%) and power (70%, 80%, 90% and 95%) levels. Settings allow entering corrected instead of uncorrected (default) correlations, changing the number of Monte Carlo replications, setting a seed for replicability, and changing the reporting precision. In sum, the application provides an accessible platform for BDV and MDV analysis based on SSD.

4.6 Discussion

This chapter presented a framework to assess discriminant validity within and between stages in process models. It then proposed a new multivariate criterion for discriminant validity of measures of constructs. Existing BDV criteria account for the pairwise associations between measures of constructs. The new MDV criterion takes all associations in a set of measures into account. It accounts for the possibility that a focal measure of a construct is fully accounted for by a combination of two or more other measures, while all pairs taken separately express discriminant validity.

A literature review of 23 multiple mediation studies in marketing found that BDV was rarely assessed (only in 5 out of 15 articles), and MDV was unsurprisingly not reported on. If discriminant validity was assessed, it was uncommon that statistical evidence was reported. However, a meta-analysis of these studies found moderate to high multiple correlations between the measures of the constructs (a median corrected correlation of about .70, the maximum was a high .92). Four follow-up case studies demonstrated the importance of assessing MDV and revealed situations where MDV was not met despite strong support for BDV. These results challenge the meaningfulness of the proposed parallel mediation models. Finally, an online Shiny application available at https://github.com/constantpieters/dv facilitates the accessibility of establishing BDV and MDV.

Of course, violations of discriminant validity can be anticipated and prevented. In the study planning phase, clear concept definitions (Podsakoff et al. 2016) and the use of measures that operationalize focal constructs but not tap into related constructs can prevent discriminant invalidity. Furthermore, reliable measurement, large sample sizes (Pieters 2017), and avoiding inflated inter-construct variances due to common method variance (MacKenzie and Podsakoff 2012; Pieters 2017) might aid in establishing discriminant validity.

Validation of measures of constructs is basic theory testing (Smith 2005), and lack of empirical support for discriminant validity does not to corroborate the theory of construct distinctiveness. Although constructs might be semantically different from other related constructs, their measures are at risk of being empirically indistinguishable, putting them at risk of redundancy. Several domains in marketing research are at risk of not attaining discriminant validity. Multidimensional measures such as need for uniqueness (Tian et al. 2001), materialism (Richins and Dawson 1992), and market-orientation (Narver and Slater 1990) can fail to attain discriminant validity if their measures overlap. Moreover, simple and multiple mediation analyses in particular should pay attention to discriminant validity. Input variables and mediators, as well as sequential mediators with each other, are by definition hypothesized to be related. Yet if inputs and mediators are indistinct, the mediator might be a manipulation check that fails to identify the purported mechanism. Mediators in parallel may also be highly correlated because they are both a function of the input(s), and can fail to express discriminant validity if they do not capture distinct processes. If measures for the mediator(s) and outcome(s) are indistinct, measures for proposed mediators could reflect measures for the outcome. In sum, lack of discriminant validity casts doubt on causal chains of process variables, and is therefore an important precondition for meaningful identification of indirect effects, and process analysis in general (Pieters 2017; Spencer et al. 2005).

Lack of discriminant validity implies that a different model is likely to better account for the data. In such cases a parsimonious theory is preferred over a broader or more general one, and generality and parsimony are important criteria to evaluate theories (Gawronski and Bodenhausen 2015). For example, dropping one of three measures of the multidimensional market orientation construct (e.g., customer orientation) would be inconsistent with the purported theory of market orientation consisting of three subdimensions (Narver and Slater 1990). Measures can be combined in a common factor or a higher order construct (Kalnins 2018). Specifically, lack of discriminant validity in process models suggests a single-process model or a model with a direct effect of the treatment on two measures of a single outcome. For instance, a single morality construct may be best represented by the measures of care and fairness in Case Study 3 due to their high correlation. Yet, dropping measures that cover subsets of the theoretical domain can lead to new validity issues if the remaining measures not account for the entire construct or lead to omitted variable bias. More generally, modifying theories and measures based on data inspection might be questionable (Gelman and Loken 2014; Simmons et al. 2011).

In sum, this chapter presented a framework to assess discriminant validity within and between stages in process models. Measures of constructs with high correlations between them are at risk of not expressing discriminant validity, and this chapter presented a new multivariate criterion that takes the full set of correlations between measures into account, instead of relying on pairwise bivariate tests. The intention of this work is to stimulate researchers to be on their guard against bivariate and multivariate discriminant invalidity as a threat to construct validity. We hope that our online application makes assessing bivariate and multivariate discriminant validity more accessible.

Appendix of Chapter 4: Details of the Shiny Application.

The application can be found at https://github.com/constantpieters/dv. Four steps determine the power of discriminant validity: 1) Input, 2) Settings, 3) Run Analysis, 4) Inspect Results. The figures explain each setting and output element in more detail.

Figure A4.1

Usage of the App: Step 1 - Input

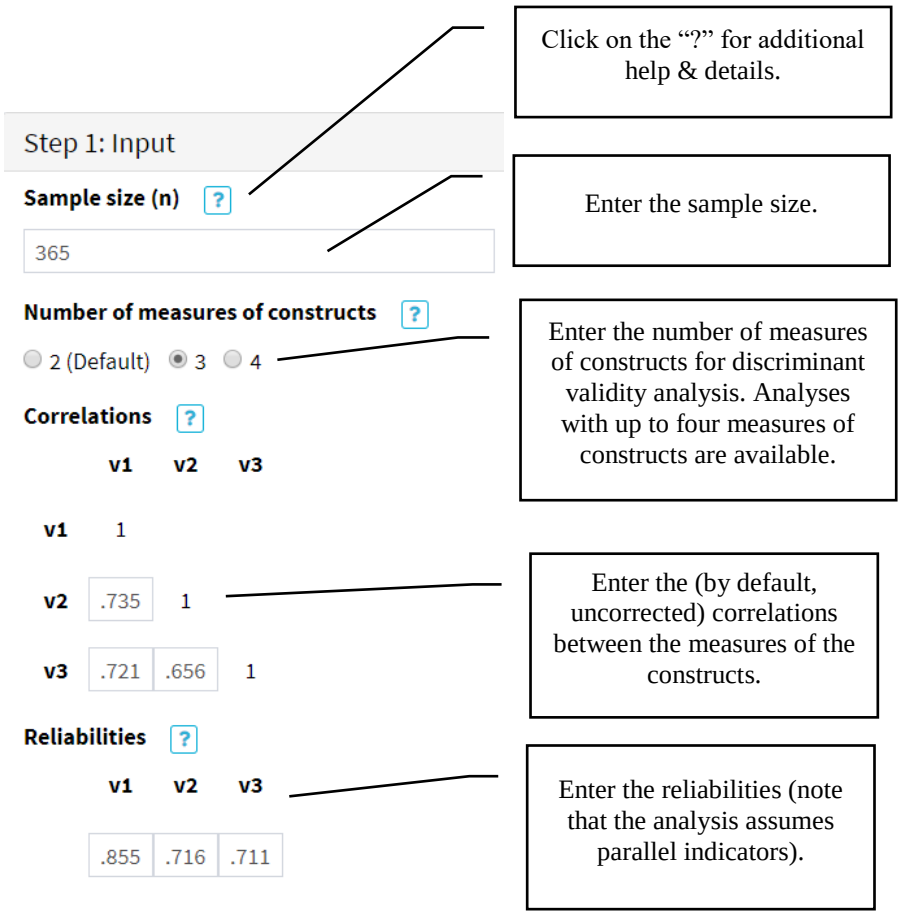


Figure A4.2

Usage of the App: Step 2 - Settings

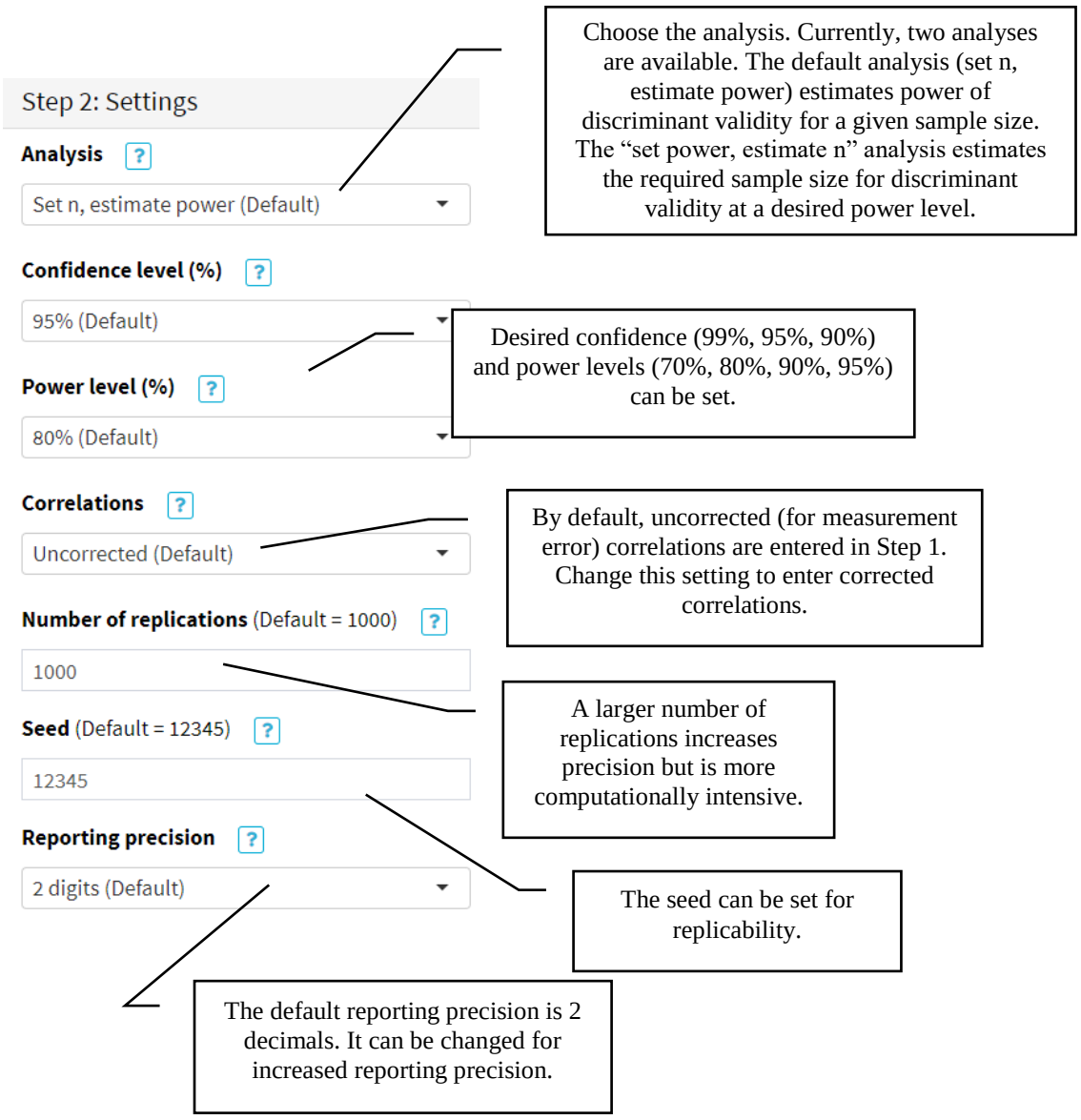
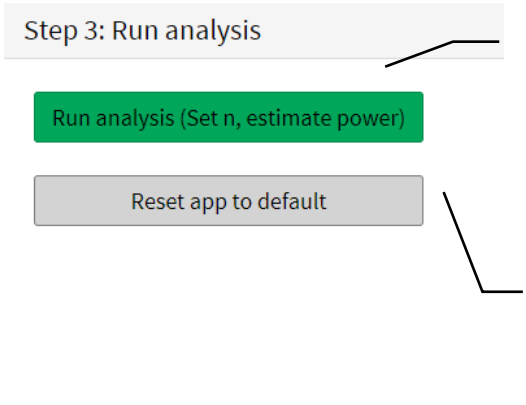


Figure A4.3

Usage of the App: Step 3 – Run Analysis

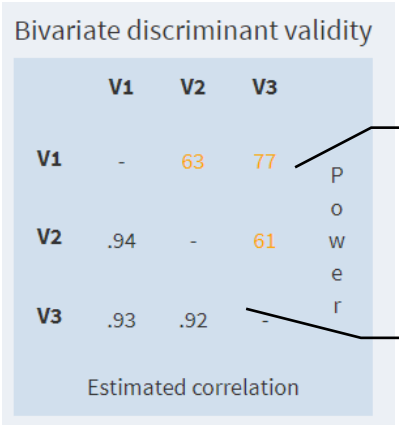


Press the button to run the analysis. The app shows a progress bar which displays the simulation progress.

The second button resets the app.

Figure A4.4

Step 4 – Inspect Output: Bivariate Discriminant Validity

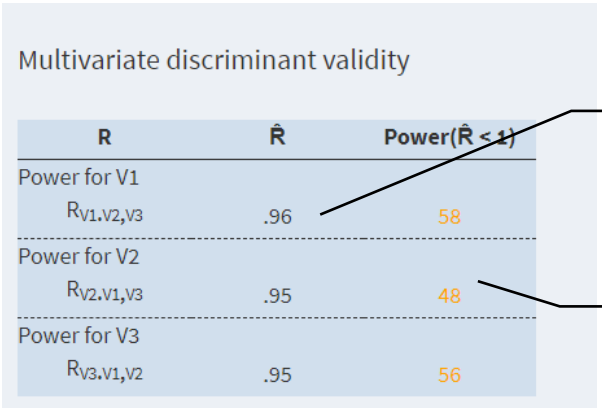


Elements above the diagonal report the estimated statistical power (in %) of the discriminant validity criterion (here: correlation-based with T = 1).

Elements below the diagonal report the estimated correlations (across the Monte Carlo replications).

Figure A4.5

Step 4 – Inspect Output: Multivariate Discriminant Validity

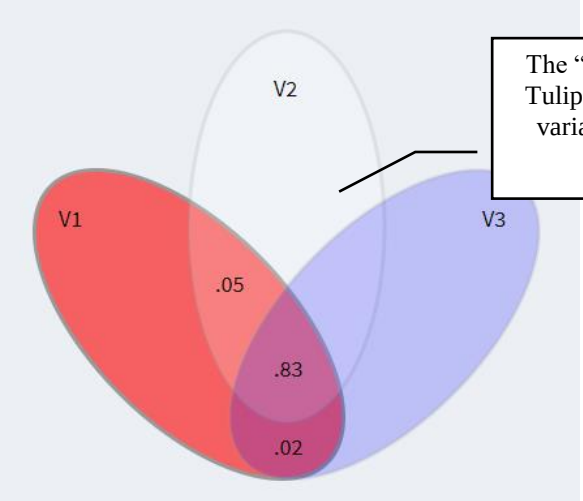


The second column shows the estimated multiple correlation (across the Monte Carlo replications).

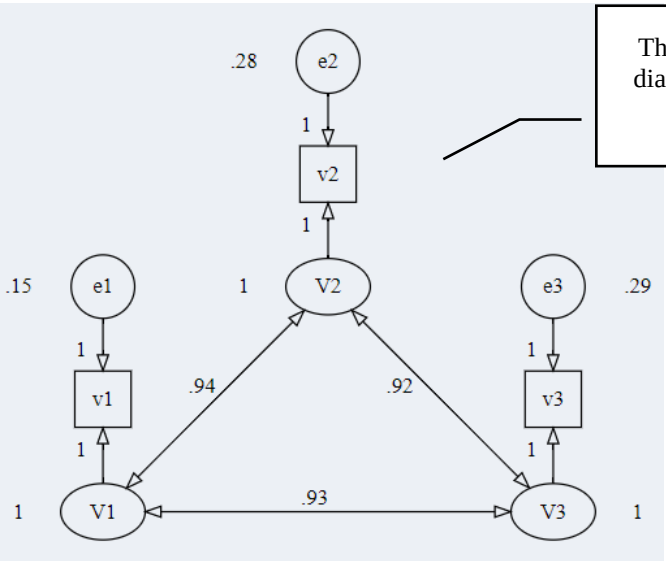
The third column contains the estimated power of the (here: correlation-based) multivariate discriminant validity criterion (across the Monte Carlo replications)

Figure A4.6

Details: Multivariate Discriminant Validity



The "Details" tab of the output has a dynamic Tulipogram. When hovered over, it shows the variance contributions to the target measure (here of V_1 with V_2 and V_3).



The estimation model used is visualized in a diagram with estimated parameters (averages across Monte Carlo replications).

Chapter 5 – General Discussion

This chapter summarizes and has follow-up analyses that provide additional insights and address issues that remain. The concluding section discusses the road ahead.

5.1 Summary

Process analysis was developed to explain fur color and birth weight of generations of guinea pigs (Wright 1921; Wright 1920). It gives insights in the mediators (how) and moderators (when) that drive the effects of input variables on relevant outcomes and has become an indispensable tool for contemporary marketing research. This dissertation presented three essays on process analysis. Table 5.1 summarizes.

Chapter 2 applied mediation methods to a substantive question. It established the referral reinforcement effect: customers were more inclined to refer a product or service if it was referred to them. Four studies, a large-scale field experiment among ridesharing customers, a reanalysis of published data from a bank's referral program, a new survey among moviegoers, and a controlled experiment, provided evidence for the referral reinforcement effect in different settings, for incentivized and organic referrals, and using different methods. Consistent with prior research, mediation analyses revealed that satisfaction accounted for about 40% of the total reinforcement effect. Yet more importantly, the remaining 60% was the non-satisfaction-mediated referral reinforcement effect. The final study explored referral motives that drive the referral reinforcement effect with a survey on customer lay beliefs. It concluded that being referred amplifies other-directed motives, such as concerns for others, that motivate referred customers to refer in turn. These results contribute to the customer engagement literature, the literature on word-of-mouth motivations, and are relevant for managers. The results are good news for managers who aim to grow their customer base with referrals. Referrals lead to more referrals through an increased satisfaction. However, managers should not only focus on their most satisfied

customers. Being referred increases a customer's inclination to refer in turn, and the majority of this effect circumvents satisfaction.

Chapter 3 compared six existing moderation methods in the face of measurement error. About 89% of 504 moderation analyses published in *Journal of Marketing* and *Journal of Marketing Research* during 2000–2017 used the means and multi-group methods that do not adequately account for measurement error to estimate the moderation effect. Monte Carlo simulations revealed that even when the interacting variables have high reliabilities of .80, the moderation effects estimated by these two methods were biased downward by more than 30%. Four other methods—factor scores, corrected means, product indicators, and latent product—account for measurement error in different ways. The latent product method was least biased and had the highest statistical power. The factor scores method had comparable performance to the latent product method. Its simplicity might favor its use over the more complex and computationally intensive latent product method.

Chapter 4 is an attempt to extend existing discriminant validity criteria. It presented a framework to assess discriminant validity, with applications to process models. It proposed a new multivariate criterion for discriminant validity of measures of constructs. The new multivariate criteria takes all associations in a set of measures into account. Chapter 4 explored sets of up to four constructs. The multivariate criterion accounts for the possibility that a focal measure of a construct is fully accounted for by a combination of two or more other measures, while all pairs taken separately express discriminant validity. It complements existing bivariate criteria that are limited to pairs of measures in a set. Reanalyses of cases taken out of a literature review and meta-analysis of 23 recent multiple mediation studies in marketing challenged the meaningfulness of the purported multiple mediation models. An online application was developed to make discriminant validity assessment more accessible.

Table 5.1
Summary of this Dissertation: Three Essays on Process Analysis for Marketing Research

Chapter	Topic	Data source(s)	Methodology	Key findings
2	The referral reinforcement effect (mediation)	Study 1: Field experimental data from customers of a ridesharing platform (n = 200,098 customers) Study 2a: Published summary statistics data of a bank's referral program (n = 470 customers) Study 2b: Survey among movie-goers (n = 851 participants) Study 3: Lab experiment (n = 87 participants) Study 4: Survey of lay beliefs (n = 1,210 participants)	Path analyses and structural equation models with and without mediation	• Referral reinforcement effect: referred customers refer more • The mediation by customer satisfaction accounts for about 40% of this effect • Importantly, the referral reinforcement effect, unmediated by satisfaction, accounts for the remaining 60%
3	Moderation in the face of measurement error	Literature review of moderation tests with at least three indicators for the interacting constructs (n = 504 moderation tests in 97 articles published in *JM* and *JMR* between 2000 and 2017)	Meta-analysis Monte Carlo simulations Meta-ANOVA	• Only 11% of the investigated moderation tests account for measurement error • Even when reliabilities are .80, not accounting for measurement error biases moderation estimates downward > 30% • The latent product method performs best in terms of bias and statistical power • The factor scores method is an accessible and easy to use alternative
4	Construct validity in process models: discriminant validity of measures of constructs	Literature review of multiple mediation studies (n = 23 studies in 15 articles published in *JM*, *JMR* and *JCR* between 2017 and 2019)	Meta-analysis Monte Carlo simulations Online Shiny application	• Multivariate discriminant validity accounts for two measures of constructs that perfectly account for a third focal measure, whereas all pairs are bivariate discriminant valid • Measures in the investigated multiple mediation studies had high multiple correlations, up to R = .92 • Three out of four follow-up multiple mediation case studies cast doubt on the validity of the purported multiple mediation due to lack of multivariate discriminant validity, despite establishing bivariate discriminant validity • The Shiny application makes discriminant validity methods more accessible

Table 5.1 (CONTINUED)

5	Follow-up Study 1: discriminant validity in Study 2a and 2b of Chapter 2	Summary statistics data (SSD) from Appendix 2B and 2C	Monte Carlo simulations	• Strong support for bivariate and multivariate discriminant validity as preconditions for meaningful mediation analyses
	Follow-up Study 2: generalizations of moderation methods (Chapter 3) Study 2a: Single-indicators Study 2b: Non-normality Study 2c: U-shapes	Literature review from Chapter 3 Estimates of univariate non-normality in the factor scores of the multi-item constructs in Chapters 2 and 4	Monte Carlo simulations	Follow-up Study 2a: Single-indicators • Single-indicators taken from a multi-indicator scale have low levels of reliability, even if the multi-indicator scale has good reliability • The corrected single-indicator method recovers the moderation effect with limited bias, but large sample sizes (about n > 500) are required even if the multi-indicator scale has a reliability of .80 for adequate power Follow-up Study 2b: Non-normality • Product terms are non-normally distributed, even when the components are normally distributed • Product terms exacerbate non-normality in the components • The factor scores and latent product methods are robust against investigated levels of non-normality in the true scores Follow-up Study 2c: U-shapes • Squared terms have the same reliability as interaction terms when components are uncorrelated, yet have lower reliability when components are correlated • When components are correlated and measurement error is unaccounted for, the bias of interaction effects is smaller than the bias of U-shaped effects • Effects of squared terms have higher statistical power than interaction effects
	Follow-up Study 3: investigating multicollinearity between discriminant valid measures of constructs (Chapter 4)		Monte Carlo simulations	• Low to moderate correlations between measures of constructs inflate standard errors and lower statistical power • Low to moderate correlations between measures of constructs do not lead to estimation bias

Notes: n refers to the sample size. *JM* is the *Journal of Marketing*, *JMR* the *Journal of Marketing Research*, and *JCR* is the *Journal of Consumer Research.*

Overall, these three Chapters contribute to meaningful and valid process modeling in marketing research. This dissertation investigated the *how* (mediation; Chapters 2 and 4), *when* (moderation; Chapter 3), and the distinctiveness of measures of constructs as a precondition for making inferences on the *how* and *when* with process analyses (Chapter 4).

Nevertheless, several issues remain. The remainder of the current Chapter 5 addresses several remaining issues with follow-up analyses. It first assesses multivariate discriminant validity in Chapter 2. A second analysis explores generalizations of moderation methods and conditions in Chapter 3. Then, Chapter 4 is followed up on by investigating the extent to which multicollinearity between discriminant valid constructs can still impact results. The final section zooms out and discusses the breadth of process analysis, with implications for its future usage in marketing research.

5.2 Follow-Up Study 1: Referral Reinforcement – Discriminant Validity

Chapter 2 applied mediation methods to investigate a substantive question: to what extent is the referral reinforcement effect mediated and non-mediated by satisfaction? Studies 2a (retail banking) and 2b (movies) investigated and quantified satisfaction-mediated and non-satisfaction-mediated referral reinforcement effects. Although both analyses expressed bivariate discriminant validity (BDV), they did not investigate multivariate discriminant validity (MDV), explored in Chapter 4. This follow-up study assesses within and between stage MDV of the measures in the mediation analyses of Studies 2a and 2b. Chapter 4 has details on the method, and the discussion here focuses on the results.

5.2.1 Chapter 2 – Study 2a (Retail banking).

Table 5.2 shows the results for Study 2a. The disattenuated correlations are moderate (highest r = .62 between SAT and REFERRING). Both BDV criteria are met, with 100% power. Similarly, there was strong evidence for the correlation- and reliability-based MDV criteria,

with 100% estimated power throughout (highest R = .65 for REFERRING). In sum, Study 2a has strong support for discriminant validity as a precondition for mediation analysis.

5.2.2 Chapter 2 – Study 2b (Movies).

Table 5.3 has the results for Study 2b. Similar to the results for Study 2a, strong support for BDV and MDV is found. The analyses assumed a reliability of 1 for the REFERRED and REFERRING single-indicators. Although this is reasonable for concrete and unidimensional constructs and measures such as self-reported referrals (Bergkvist and Rossiter 2007), a sensitivity analysis lowered the assumed single-indicator reliability of 1 in steps of .05 until 80% or lower power was reached. Single-indicator reliabilities lower than .80 did not attain reliability-based MDV ($R_{REFERRING}$ = .75, power = 53%) and reliabilities < .75 failed BDV ($r_{SAT,REFERRING}$ = .68, power = 64%). Single-indicator reliabilities lower than .55 resulted in correlation-based MDV criteria not being met ($R_{REFERRING}$ = .95, power = 34%). In sum, although it is reasonable to expect that the single-indicator referral measures have high reliabilities due to their concreteness, discriminant validity is supported for a wide range of reliabilities.

5.3 Follow-Up Study 2: Moderation – Generalizations

Chapter 3 investigated six moderation methods in the face of measurement error. A large-scale Monte Carlo simulation study showed that the latent product method performed best, but that factor scores are an easy-to-use alternative with comparable performance. Although the simulations had population parameters that were based on the results of an extensive literature review, it had several restrictions. A follow-up study in Chapter 3 relaxed the restriction that all indicators were equally good. Three additional generalizations are investigated here: single-indicator measurement (on the level of the indicators), non-normality (in latent variables), and U-shapes (on the level of the structural model).

Table 5.2
Follow-Up Analysis of Bivariate and Multivariate Discriminant Validity in Study 2a of Chapter 2

Panel A: Summary Statistics Data (SSD)

		X REFERRED	M SAT	Y REFERRING
X	REFERRED	(1 / 1)		
M	SAT	.18	(2 / .85)	
Y	REFERRING	.28	.56	(5 / .94)

Panel B: Bivariate Discriminant Validity (BDV) - Correlation-Based Criterion

		X REFERRED	M SAT	Y REFERRING
X	REFERRED	1	100%	100%
M	SAT	.19 [.10, .29]	1	100%
Y	REFERRING	.29 [.02, .38]	.62 [.56, .69]	1

Panel C: Bivariate Discriminant Validity (BDV) - Reliability-Based Criterion

		X REFERRED	M SAT	Y REFERRING
X	REFERRED	(1)	100%; 100%	100%; 100%
M	SAT	.19 [.01, .29]	(.85) [.82, .88]	100%; 100%
Y	REFERRING	.29 [.02, .38]	.62 [.56, .69]	(.94) [.93, .95]

Panel D: Multivariate Discriminant Validity (MDV) – Correlation- & Reliability-Based Criteria

		Estimated multiple R	Estimated reliability ($r_{Vi,Vi}$)	% Power correlation-based criterion (1 – R)	% Power reliability-based criterion ($r_{Vi,Vi}$ - R)
X	REFERRED	.30 [.21, .38]	1	100%	100%
M	SAT	.63 [.56, .69]	.85 [.92, .88]	100%	100%
Y	REFERRING	.65 [.59, .71]	.94 [.93, .95]	100%	100%

Notes: n = 470. Estimates are based on Monte Carlo simulations with 1,000 replications. Mean 95% confidence intervals are between brackets. Panel A has bivariate uncorrected correlations between the measures of the constructs and # indicators / reliability in parentheses on the diagonal, for details see Appendix 2B. Panel B has estimated bivariate corrected correlations below the diagonal and power of the correlation-based criterion with T = 1 above the diagonal. Panel C has reliabilities on the diagonal in parentheses and power of the reliability-based criterion (T = $r_{Vi,Vi}$) with respect to the measures in the row and column (separated by a ;) above the diagonal.

Table 5.3
Follow-Up Analysis of Bivariate and Multivariate Discriminant Validity in Study 2b of Chapter 2

Panel A: Summary Statistics Data (SSD)

		X REFERRED	M SAT	Y REFERRING
X	REFERRED	(1 / 1)		
M	SAT	.16	(3 / .77)	
Y	REFERRING	.38	.52	(1 / 1)

Panel B: Bivariate Discriminant Validity (BDV) - Correlation-Based Criterion

		X REFERRED	M SAT	Y REFERRING
X	REFERRED	1	100%	100%
M	SAT	.18 [.11, .26]	1	100%
Y	REFERRING	.38 [.33, .44]	.59 [.54, .64]	1

Panel C: Bivariate Discriminant Validity (BDV) - Reliability-Based Criterion

		X REFERRED	M SAT	Y REFERRING
X	REFERRED	(1)	100%; 100%	100%; 100%
M	SAT	.18 [.11, .26]	(.77) [.74, .80]	100%; 100%
Y	REFERRING	.38 [.33, .44]	.59 [.54, .64]	(1)

Panel D: Multivariate Discriminant Validity (MDV) – Correlation- & Reliability-Based Criteria

		Estimated multiple R	Estimated reliability ($r_{Vi,Vi}$)	% Power correlation-based criterion (1 – R)	% Power reliability-based criterion ($r_{Vi,Vi}$ - R)
X	REFERRED	.39 [.33, .45]	1	100%	100%
M	SAT	.59 [.54, .65]	.77 [.74, .80]	100%	100%
Y	REFERRING	.65 [.61, .70]	1	100%	100%

Notes: n = 851. Estimates are based on Monte Carlo simulations with 1,000 replications. Mean 95% confidence intervals are between brackets. Panel A has bivariate uncorrected correlations between the measures of the constructs and # indicators / reliability in parentheses on the diagonal, for details see Appendix 2C. Panel B has estimated bivariate corrected correlations below the diagonal and power of the correlation-based criterion with T = 1 above the diagonal. Panel C has reliabilities on the diagonal in parentheses and power of the reliability-based criterion (T = $r_{Vi,Vi}$) with respect to the measures in the row and column (separated by a ;) above the diagonal.

5.3.1 Study 2a – Single-indicators.

Single-indicators are common in marketing process analyses. Pieters (2017) found that out of 166 mediation analyses 86 articles using experiments published in the *Journal of Consumer Research* between 2014-2016, 43% of mediators and 64% of outcomes were measured with single-indicators. For instance, Ma and Roese (2014) find in Study 1b (n = 62 MTurk participants) that a maximizing mindset (X, manipulated), had a positive effect on the likelihood of returning a smartphone (Y, measured with a single-indicator), mediated by regret (M_1, measured with a single-indicator) as well as satisfaction (M_2, measured with a single-indicator) in parallel.

Moderation studies in experimental and marketing strategy research might also include single-indicators. Chapter 3 focused on multi-indicator explanatory latent variables. Yet, among the 504 investigated moderation effects in Chapter 3, 147 (29%) had a single-indicator Y. As an example of single-indicator explanatory variables, Homburg and Bucerius (2005) found across 232 mergers and acquisitions that the positive relationship between the speed of integration (X, 8 indicators, reliability = .89) on performance (Y, 2 indicators, reliability = .76) was stronger for mergers and acquisitions that had a higher relative size of the acquired firm (Z, measured with a single-indicator).

Reliability of single-indicator measures.

Single-indicator measures can be appropriate for concrete unidimensional constructs. They decrease questionnaire length, respondent fatigue, and avoid variance due to common methods within a scale. Yet, they might have decreased coverage, reliability and validity of abstract or multidimensional constructs (Bergkvist and Rossiter 2007; Petrescu 2013; Pieters 2017). The discussion here focuses on reliability. The Spearman-Brown prophecy formula estimates the reliability of a single-indicator measure if it is taken from a multi-indicator

measure with known reliability, under the assumption that each indicator in the larger scale is equally good. It is:

$$r_{Vi,Vi}^{single} = \frac{r_{Vi,Vi}}{r_{Vi,Vi} + u(1 - r_{Vi,Vi})}, \quad (5.1)$$

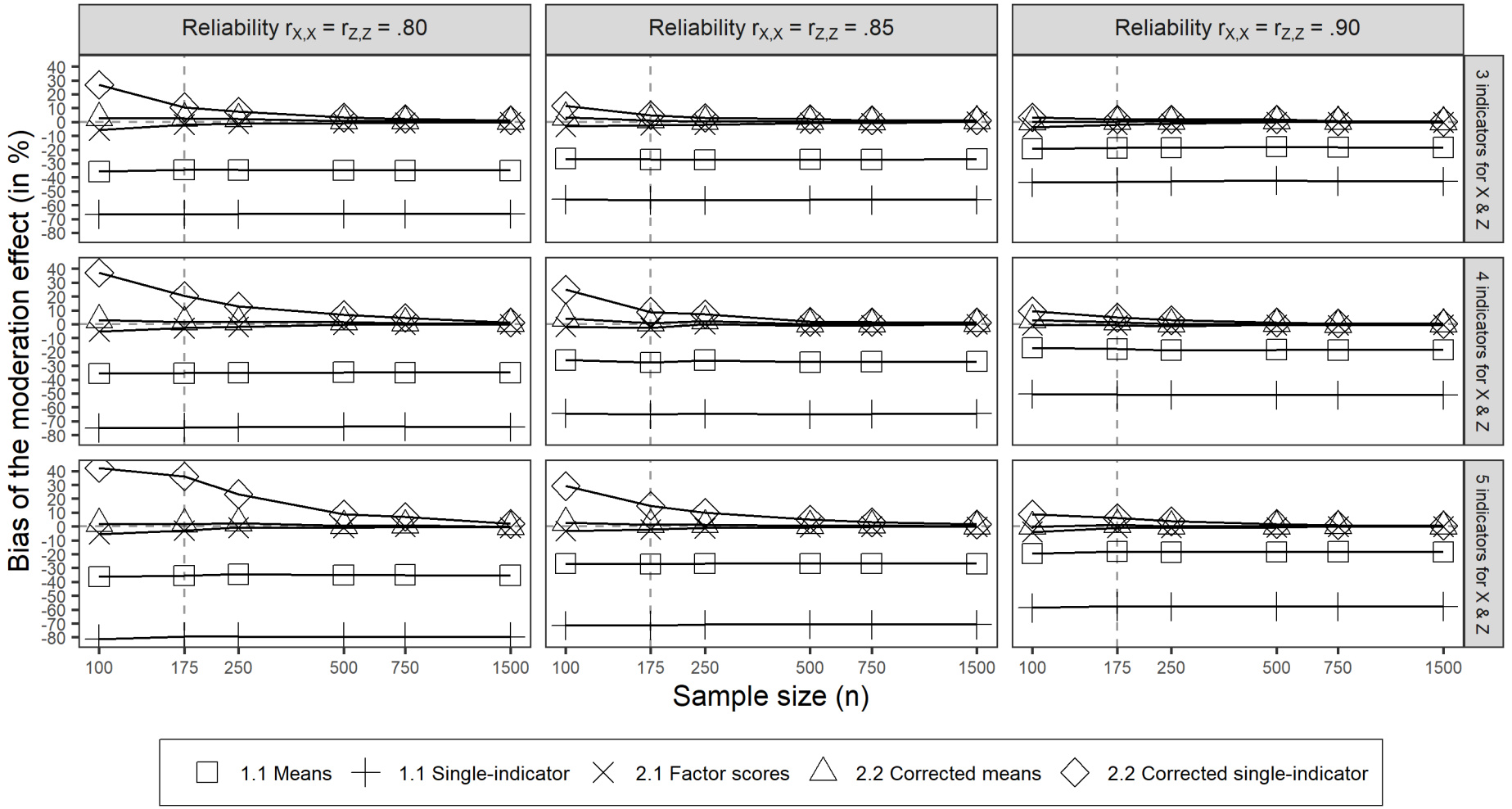
where $r_{Vi,Vi}^{single}$ is the estimated single-indicator reliability, $r_{Vi,Vi}$ the provided reliability of the multi-indicator scale, and u the number of indicators of the measure with multiple indicators (Pieters 2017). For instance, if a four-indicator measure with a good to excellent reliability of .85 (about the median values in the literature review of Chapter 3) is reduced to a single-indicator, its reliability becomes .59, which is commonly regarded as inadequate (Peterson 1994). This reliability becomes even lower if the original scale becomes larger, for instance if the single-indicator is from a five-indicator scale, resulting in a reliability of .53.

Thus, it is expected that using single-indicators without accounting for their unreliability can lead to severely biased moderation effects, even if they are taken from reliable multi-indicator scales. The bias can increase when single-indicators are from unreliable and large multi-indicator scales (cf. Equation 5.1). Moderation estimates using single-indicators might also have low power due to multiplication of low reliability indicators. Yet, it is unclear to what extent single-indicators are able to attain the level of power of their original scale. Moreover, the question remains which sample size attains acceptable levels of statistical power. Monte Carlo simulations further investigate this.

Monte Carlo simulations.

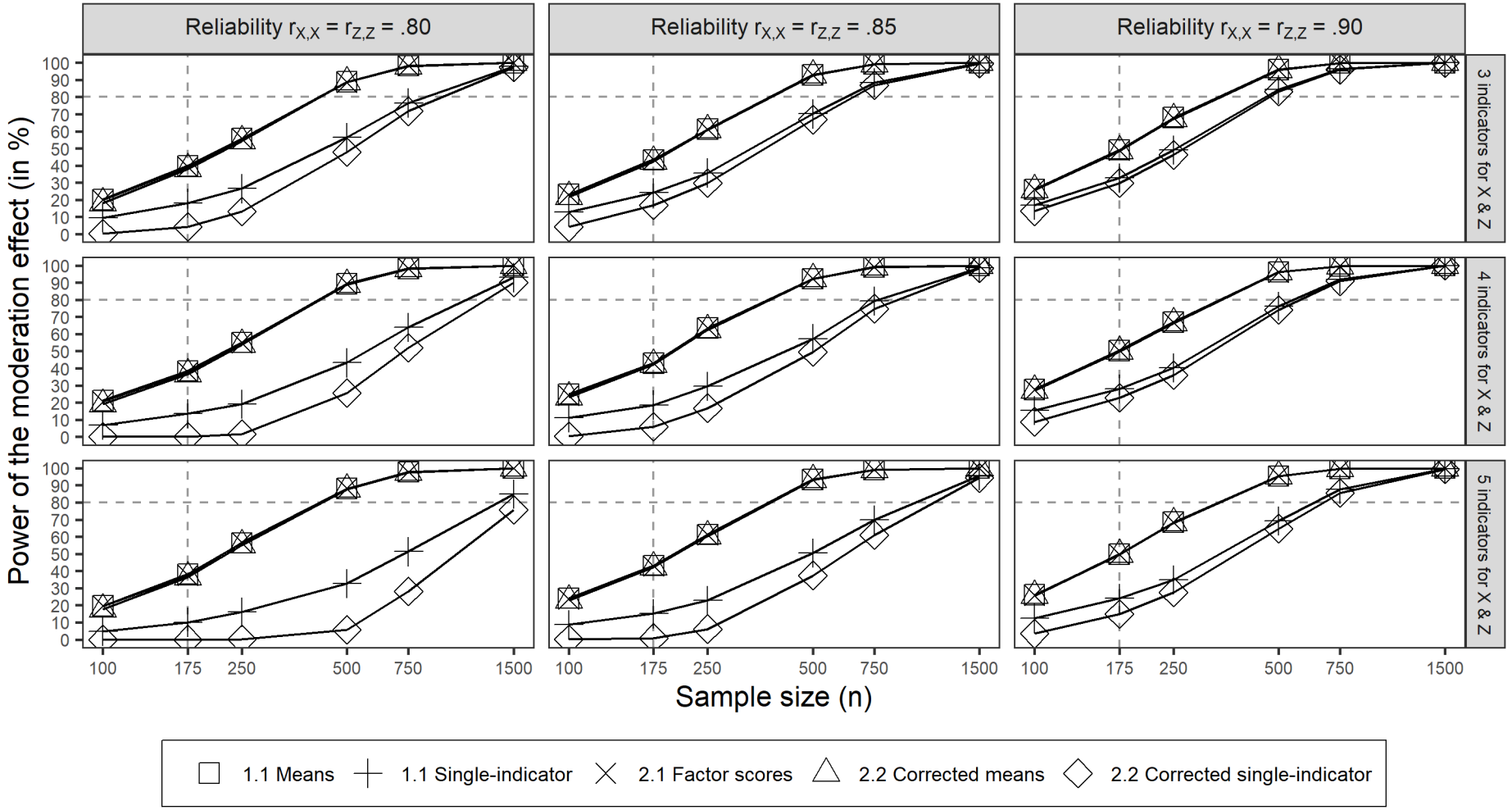
Follow-up Monte Carlo simulations with 5,000 replications per cell investigated the bias and power of single-indicator measures. The design varied the sample sizes from 100 to 1,500 with effect sizes of the main and moderation effects of .20, and a correlation between X and Z of .20. Unlike Chapter 3, which had 3 indicators for X and Z, the follow-up simulation fixed the number of indicators to 3, 4 (the median in the literature review) or 5.

Figure 5.1
Bias of the Moderation Effect for Single- and Multi-Indicator Scales



Notes: Results are based on a Monte Carlo simulation with 5,000 replications. Panels visualize the estimated bias (in %) of the moderation effect across sample sizes (scale is log-transformed) for different reliabilities and number of indicators of X and Z. Dashed lines indicate a sample size of 175 and zero bias.

Figure 5.2
Power of the Moderation Effect for Single- and Multi-Indicator Scales



Notes: Results are based on a Monte Carlo simulation with 5,000 replications. Panels visualize the estimated power (in %) of the moderation effect across sample sizes (scale is log-transformed) for different reliabilities and number of indicators of X and Z. Dashed lines indicate a sample size of 175 and 80% power.

The reliability of X and Z was fixed to .80, .85 (about the median in the literature review) or .90. Additional details of the method and population model are in Chapter 3.

Five methods estimated the moderation effect. Methods 1.1 (Means), 2.1 (Factor scores) and 2.2 (Corrected means) are included as benchmark methods. The simulation also estimated two single-indicator models using randomly selected indicators, one in X and one in Z, in each replication. Method 1.1 (Single-indicator) uses a single-indicator to estimate the moderation effect. Method 2.2 (Corrected single-indicator) corrects the single-indicator for measurement error. It uses the known multi-indicator reliabilities (.80 or .90) and Equation 5.1 to estimate the single-indicator reliabilities, which are then entered in Equation 3.4 to estimate the reliability of the single-product-indicator.

Figure 5.1 visualizes the estimated bias of the moderation effect across sample sizes, reliabilities, and number of indicators in the multi-indicator scale. First, the results for Methods 1.1 (Means), 2.1 (Factor scores) and 2.2 (Corrected means), using the multi-indicator scale, parallel those in Chapter 3. Using means is biased, for about -35% when $r_{X,X}$ and $r_{Z,Z}$ are .80, and factor scores and corrected means are able to recover the moderation effect with little bias.

Second, for the single-indicators, Method 1.1 (Single-indicator) is substantially biased, e.g., a downward bias of about -65% for the median reliability of .85 and 4 indicators of X and Z. This bias is more than double the bias of -27% when using the multi-indicator scale. Third, accounting for measurement error in the single-indicator recovers the true moderation effect, yet it might overestimate the moderation effect when samples are small and single-indicators are taken from larger multi-indicator scales. It supports the efficacy of the corrected means method when single-indicators are used.

Figure 5.2 visualizes the power of the moderation effect. First, using multi-indicator scales, Methods 1.1 (Means), 2.1 (Factor scores) and 2.2 (Corrected means), have similar

levels of statistical power, although at the median sample size of about 175, 80% power is not attained in any of the investigated conditions. Second, the low reliability of single-indicators leads to low power of the moderation effect. For instance, at the median sample size of 175, and four-indicator X and Z measures with .85 reliability, the power of the (uncorrected) single-indicator Method is about 19%. Third, and most interestingly, while accounting for measurement error in the single-indicator successfully accounts for its measurement error, the power is much smaller than that of the uncorrected single- and multi-indicator methods, particularly when the number of indicators in the scale increases. Large samples are then needed to attain adequate power levels. As shown in the bottom left panel of Figure 5.2, samples larger than about 1,500 are needed for 80% power of the moderation effect when X and Z are measured with 5 indicators and have a reliability of .80.

In sum, the results show that single-indicators that are taken from multi-indicator scales are best avoided in small samples, even when the multiple indicators are equally good and the scale has good to excellent reliability. Accounting for measurement error requires large samples, about 500(700) observations for 3(5) indicators for X and Z and reliabilities of .90 for unbiased and adequately powered moderation effects. Interestingly, accounting for measurement error in a single-indicator might have substantial lower power levels than not accounting for measurement error in the respective multi-indicator scale. For instance, at a sample size of 750, a reliability of .80, and 5 indicators for X and Z, the uncorrected means method was biased downward yet had a power of about 98%, while the corrected single indicator was virtually unbiased and had a power of about 28%.

Measurement of broader or multidimensional constructs, such as need for uniqueness (Tian et al. 2001), materialism (Richins and Dawson 1992), and market-orientation (Narver and Slater 1990), requires longer scales (Flake et al. 2017). Yet, single-indicators might be feasible in large samples if accurate reliability estimates are available, for instance from pilot-

studies, meta-analyses, or specific prior research, and unreliability in the single-indicator is accounted for. They can also be used if taken from multi-indicator scales that have high reliabilities. For instance, a single-indicator from a three-indicator scale with a high reliability of .95 has a reliability of .86, which can be considered good to excellent, although the reliability of the product term could still be very low. Single-indicators are then best reserved for large samples and concrete and general constructs such as beliefs, perceptions, intentions and satisfaction (Bergkvist and Rossiter 2007). Such constructs are unidimensional and envisioned and understood identically by virtually everyone (Rossiter 2002).

5.3.2 Study 2b – Non-normality.

Non-normality of indicators is potentially widespread in marketing research, but rarely tested (Hulland et al. 1996). There are two independent sources of non-normality in indicator distributions (Hu and Bentler 1999). First, there can be non-normality in true latent variable distributions, which is then reflected in indicators that capture the latent variable. For instance, Peterson and Wilson (1992) concluded that "[v]irtually all self-reports of customer satisfaction possess a distribution in which a majority of the responses indicate that customers are satisfied and the distribution itself is negatively skewed" (p. 62). Second, the distribution of measurement errors of the indicators themselves can be non-normal if unexplained factors are distributed non-normally, even when the true scores are normally distributed. This follow-up analysis investigates non-normality in the true scores. A likely reason for the skewness in satisfaction measures is that the distributions reflect true satisfaction, and that consumers are for the most part satisfied with what they choose to purchase and consume (Peterson and Wilson 1992).

Non-normality in latent variable distributions.

The univariate skewness and excess kurtosis of a distribution are common estimates of the degree of non-normality in that distribution (Curran et al. 1996; Finch et al. 1997; Finney and

DiStefano 2006; Moosbrugger et al. 1997). Theoretically normal distributions have a skewness and excess kurtosis of zero, and deviations from zero reflect the degree of non-normality. Skewness results in asymmetry of the distribution. For example, negative skewness reflects that customers are generally satisfied and not dissatisfied with the products they purchase and consume (Peterson and Wilson 1992). A positive kurtosis reflects a higher likelihood that there are extreme observations in the tails of the distribution than there would be in a normal distribution.

Table 5.4 contains estimates of the skewness and kurtosis of the factor scores of the multi-item constructs in this dissertation. The satisfaction measure in Study 2 (moviegoers) of Chapter 2 had negative skewness of -1.66, which is consistent with Peterson and Wilson (1992) and estimates between -.16 and -2.20 across product-categories reported earlier (Westbrook 1980). It had an estimated 3.00 positive kurtosis. Thus, customers are on average more satisfied (due to the negative skewness) but are also more likely to be in the tails of the satisfaction distribution (i.e., extremely dissatisfied or satisfied; due to the positive kurtosis) than would be expected in a normally distributed satisfaction score with the same mean and standard deviation. Similar results were found for the mediators in Study 3 (controlled lab study) of Chapter 2 (skewness estimates -1.65 and -1.14 for affective and cognitive evaluation respectively; kurtosis was 3.03 and 3.37). Table 5.4 also contains estimates of the non-normality in factor scores of the multi-item mediators in the multiple mediation studies investigated in Chapter 4 for which the raw data were available. For instance, the care concern mediator in Study 4 of Goenka and Van Osselaer (2019) had an estimated skewness of -1.18 and an estimated kurtosis of 2.16.

The impact of non-normality on moderation methods.

The moderation methods that were examined in Chapter 3 generally assume multivariate normality of the indicators and measurement errors (Bollen 1989; Brandt et al. 2014). In non-

moderation models, parameter estimates are robust to non-normality, but standard errors can be biased (Finney and DiStefano 2006). For instance, Finch et al. (1997) investigated non-normality in mediation models and found that the estimation bias seldom exceeded 3%.

However, moderation models contain inherent non-normality even if the latent variables are normally distributed, and they exacerbate non-normality due to the product term (Moosbrugger et al. 1997; Schermelleh-Engel et al. 1998). To illustrate this, Panel A of Figure 5.3 illustrates the density of X (top plot), Z (middle plot) and their product XZ (bottom plot). X and Z are 1,500 observations from a standard bivariate normal distribution with $r_{X,Z}$ = .20. It shows that even though X and Z follow their theoretical normal distribution (dotted line), the product XZ is non-normally distributed (estimated skewness = 1.16 and kurtosis = 6.94 in this sample).

The non-normality of the product term is exacerbated if X and Z are non-normal themselves. For illustration, the distributions in Panels B and C were obtained by

Table 5.4
Univariate Non-Normality in the Factor Scores of the Multi-Item Constructs in this Dissertation

	M		Y		C	
Study	Skewness	Kurtosis	Skewness	Kurtosis	Skewness	Kurtosis
Chapter 2: Referral Reinforcement						
Study 2: Survey among moviegoers	-1.66	3.00	-	-	-.22	-.36
					-.32	-.04
Study 3: Lab experiment	-1.65	3.03	-.75	.35	-	-
	-1.14	3.37				
Chapter 4: Discriminant Validity						
Study 4 in Goenka and Van Osselaer (2019)	-1.18	2.16	-	-	-	-
	-.73	.72				
Study 3b in Paley et al. (2018)	-.40	-.68	-	-	-	-
	-.68	.71				
Study 4 in Steffel and Williams (2018)	.22	-.86	-	-	-	-
	.41	-.64				

Notes: M refers to mediator(s), Y to outcome, and C to covariate(s). Kurtosis refers to excess kurtosis, i.e., the deviation from the kurtosis of 3 from a true normal distribution. In Study 2 of Chapter 2, M was customer satisfaction and Cs were opinion seeking and opinion leadership. In Study 3 of Chapter 2, Ms were affective and cognitive evaluation, and Y was inclination to refer. In Study 4 of Goenka and Van Osselaer (2019), Ms were care and fairness concerns. In Study 3b of Paley et al. (2018), Ms were beliefs about negative feelings and anticipated pleasure. In Study 4 of Steffel and Williams (2018), Ms were anticipated disappointment and regret.

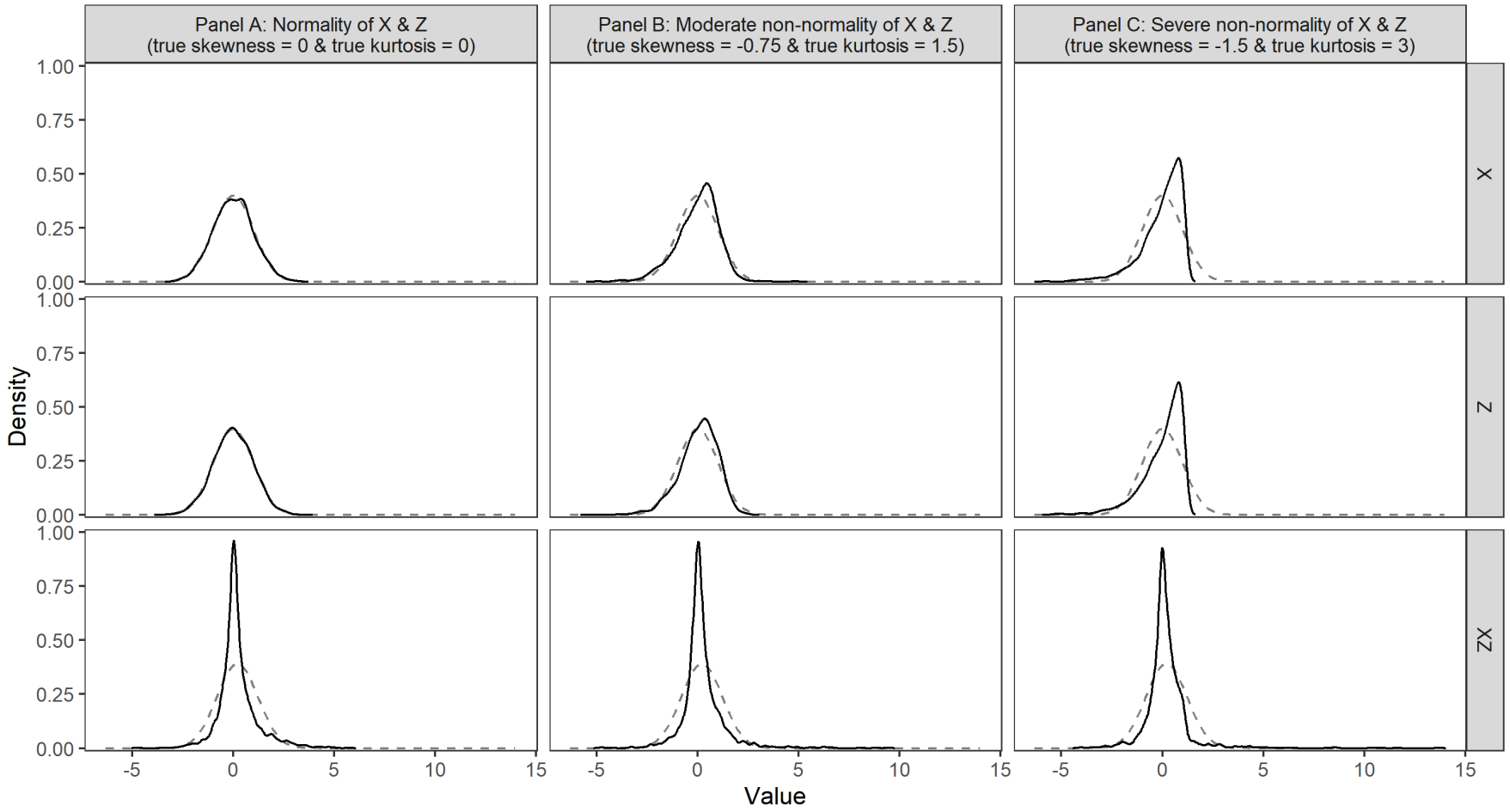
transforming the normal distributions in Panel A to obtain non-normal ones with predetermined non-zero skewness and kurtosis, using a third degree polynomial of the normally distributed data (Vale and Maurelli 1983). Skewness/kurtosis combinations of -.75/1.5 and -1.5/3 reflected moderate and severe non-normality, based on Table 5.4.[12] Panels B (moderate non-normality) and C (severe non-normality) of Figure 5.3 show that non-normality in X and Z is exacerbated by taking their product. The skewness/kurtosis levels for the product term XZ were 2.96/35.80 and 4.91/63.20 in the moderate and severely non-normality conditions.

Of course, non-normal true scores of X, Z and XZ result in non-normal true scores of Y, if there are true effects of the non-normal scores. Non-normal true scores imply non-normal indicators of X, Z and Y. In this example, the estimated skewness was -.65 and the kurtosis was 1 for the indicators of X and Z in the severe non-normality condition, assuming three indicators per factor and a reliability of .80.

Ironically, if measurement errors are normally distributed but true scores are non-normally distributed, a lower reliability attenuates the degree of non-normality of the true score that is transferred to the indicator. In sum, even when X and Z are normally distributed, their product XZ is non-normally distributed. Non-normality propagates: the outcome Y and the indicators of the non-normal scores are also non-normal. This violates multivariate normality assumptions of the methods that were examined in Chapter 3 (Bollen 1989; Brandt et al. 2014; Finney and DiStefano 2006). Nevertheless, the simulation studies that were presented in Chapter 3 showed that the preferred methods were virtually unbiased, despite the implied non-normality in XZ when X and Z are normally distributed.

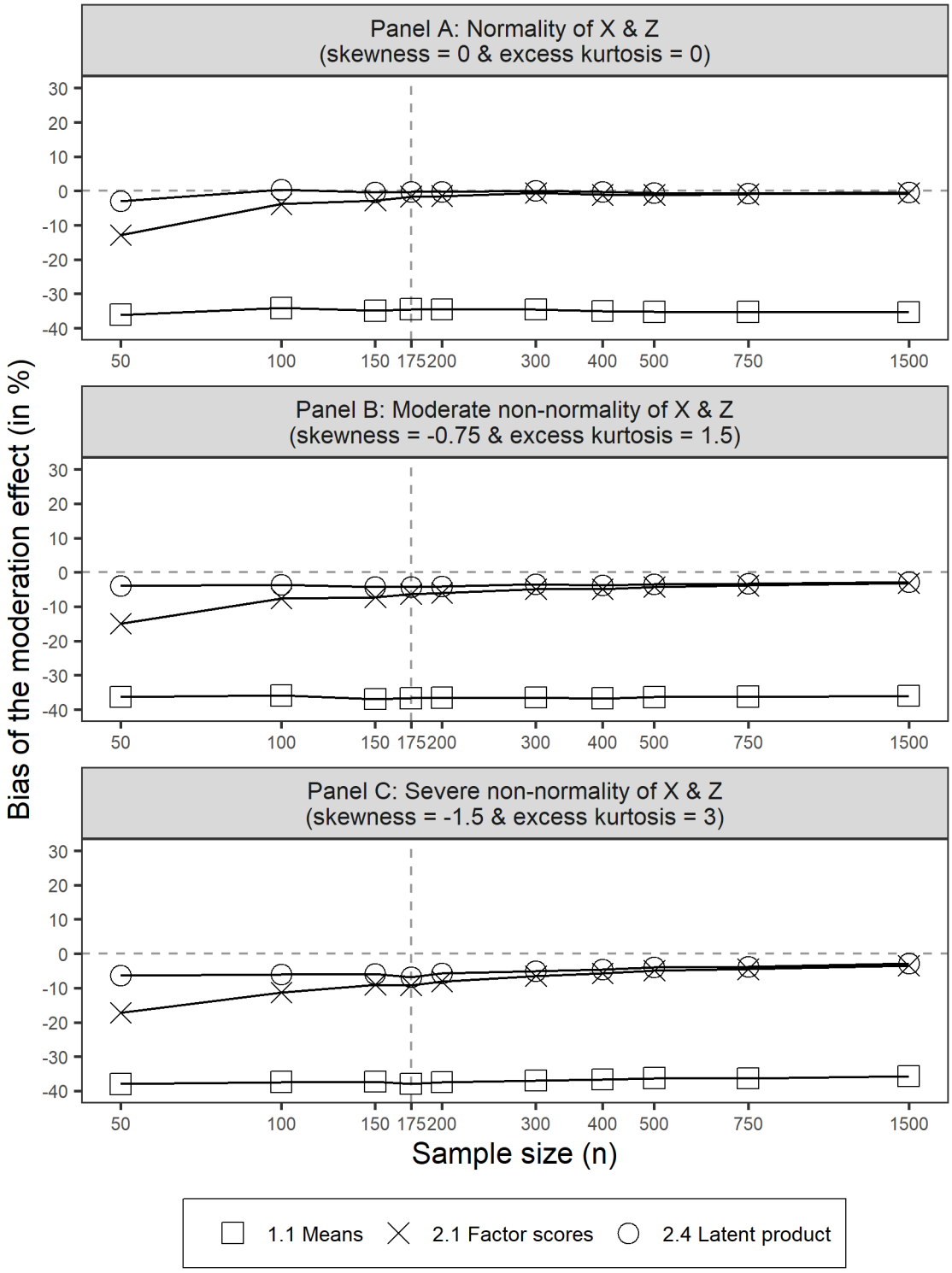
[12] There is little empirical guidance for what constitutes moderate and severe non-normality. Previous simulation studies used skewness = 2, excess kurtosis = 7 for moderate non-normality and skewness = 3, excess kurtosis = 21 for severe non-normality (Brandt et al. 2014; Curran et al. 1996; Finch et al. 1997). However, these values did not have strong empirical justification.

Figure 5.3
Non-Normality in Variables and Products



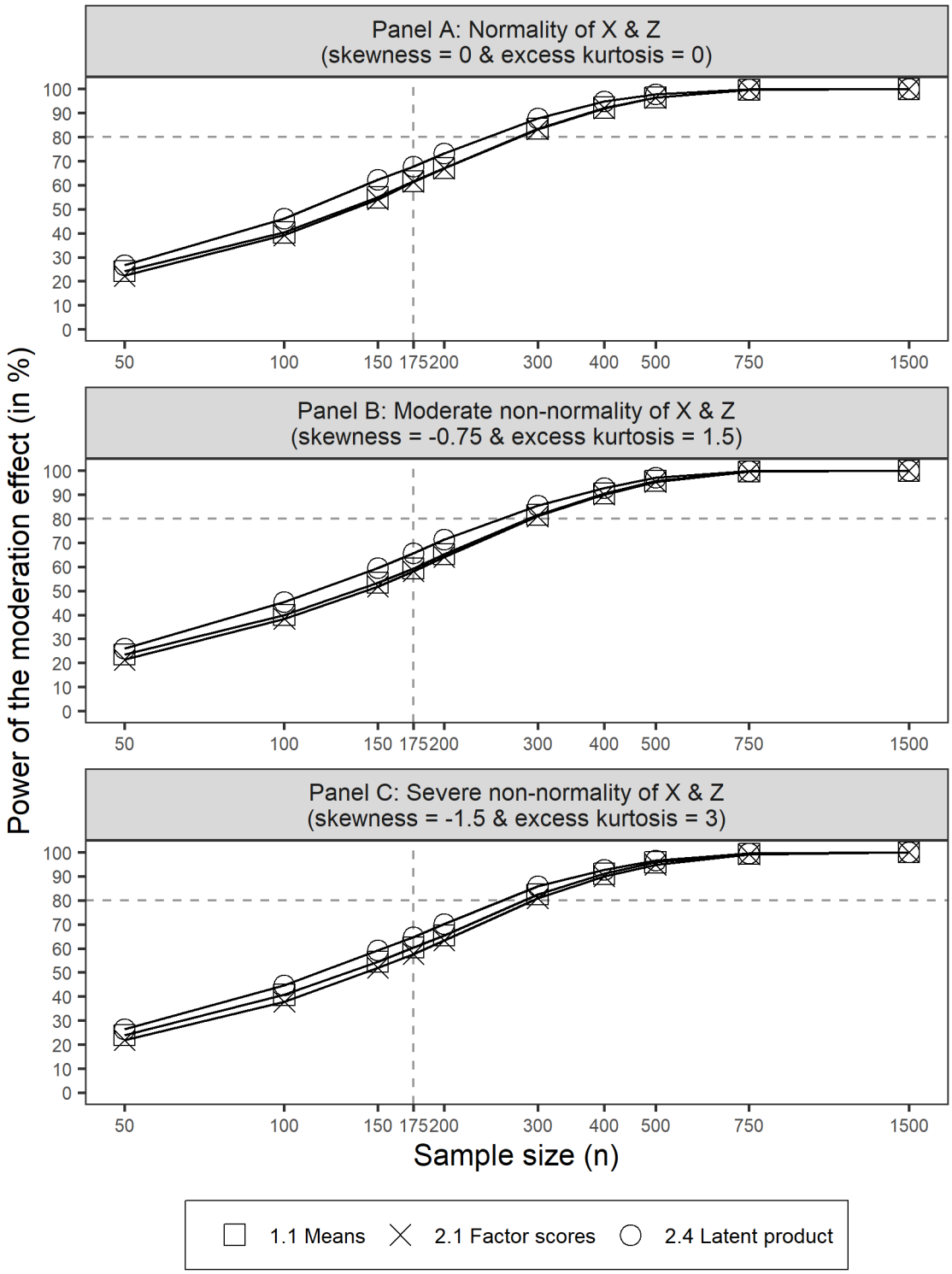
Notes: Plots contain density X, Z and their product XZ under three conditions of non-normality. Kurtosis refers to excess kurtosis, i.e., the deviation from the kurtosis of 3 from a true normal distribution. The densities (solid line) are based on 1,500 observations drawn randomly from a bivariate standard normal distributed with correlation .20 between X and Z. The dotted line is the density of a true normal distribution with the same mean and standard deviation as the (non-normal) sampled distribution (i.e., N(0,1) for X and Z and N$(.20, \sqrt{1 - .20^2})$ for XZ). The estimated skewness/excess kurtosis for XZ were 1.1/4.7, 2.2/14.3 and 4.0/34.4 respectively for normality, moderate non-normality, and severe non-normality.

Figure 5.4
Impact of Non-Normality on Bias of the Moderation Effect Across Sample Sizes



Notes: Results are based on a Monte Carlo simulation with 5,000 replications. Panels visualize the estimated bias (in %) of the moderation effect across sample sizes (scale is log-transformed) for different levels of non-normality in X and Z. Dashed lines indicate a sample size of 175 and zero bias.

Figure 5.5
Impact of Non-Normality on Power of the Moderation Effect Across Sample Sizes



Notes: Results are based on a Monte Carlo simulation with 5,000 replications. Panels visualize the estimated power (in %) of the moderation effect across sample sizes (scale is log-transformed) for different levels of non-normality in X and Z. Dashed lines indicate a sample size of 175 and 80% power.

However, the question remains to what extent the results are robust against more severe non-normality in XZ due to non-normality in X and Z. Follow-up Monte Carlo simulations investigated this.

Monte Carlo simulations.

Follow-up Monte Carlo simulations with 5,000 replications per cell investigated the impact of non-normality in X and Z on the bias and power of the preferred moderation methods. The design varied the sample sizes from 50 to 1,500 with effect sizes of the main and moderation effects of .20, and a correlation between X and Z of .20The distributions of X and Z varied: normality (skewness = 0 & kurtosis = 0), moderate non-normality (skewness = -.75 & kurtosis = 1.5) and severe non-normality (skewness = -1.5 & kurtosis = 3). As in Chapter 3, the simulation generated three indicators for X and Z each, with unity loadings and normally distributed measurement errors, and a fixed reliability of X and Z of .80. Additional details of the population model are in Chapter 3. The investigated Methods were 1.1 (Means) as a benchmark method and 2.1 (Factor scores), the preferred method in Chapter 3. The simulation also included Method 2.4 (Latent product) as it relaxes the assumption that Y can be non-normally distributed due to the non-normal product term (Kelava et al. 2011; Klein and Moosbrugger 2000).

Figure 5.4 visualizes the estimated bias of the moderation effect across sample sizes. Panel A (normality in X and Z) replicates the results from Chapter 2: Methods 1.1 (Means) is severely biased across all sample sizes, here about -36%, whereas Methods 2.1 (Factor scores) and 2.4 (Latent product) are virtually unbiased, particularly when the sample size increases. Panels B and C of Figure 5.4 introduce non-normality in X and Z. When there is moderate (Panel B) or severe (Panel C) non-normality in X and Z, Method 1.1 (Means) remains biased. However, Methods 2.1 (Factor scores) and 2.4 (Latent product) slightly underestimate the moderation effect, for about -5% to -10% for moderate sample sizes of

about 175. At large sample sizes (here: 1,500) the bias decreases and Methods 2.1 (Factor scores) and 2.4 (Latent product) underestimate the moderation effect for about 3%. This bias is minor (Muthén and Muthén 2002), and its magnitude is similar to the findings of earlier simulation studies with non-moderation methods (Finney and DiStefano 2006).Figure 5.5 has the estimated power of the moderation effect across sample sizes. Panel A plots the power for normality of X and Z. Consistent with the results in Chapter 3, Method 2.4 (Latent product) has the highest power, with only slightly lower power for Methods 1.1 (Means) and 2.1 (Factor scores). At a sample size of 175, which is about the median sample size found in the literature review in Chapter 3, no method attains 80% power as a rule of thumb for sufficient power (at best: about 68% for Latent product). Panels B and C of Figure 5.5 show that the power curves for non-normality of X and Z are virtually identical to those of the normality condition (Panel A).

It is important to note that Method 2.4 (Latent product) has stricter assumptions than Method 2.1 (Factor scores). It accounts for the non-normality in Y due to non-normality of the product term (Klein and Moosbrugger 2000). Yet, it relies on the strict assumption of normally distributed indicator distributions of X and Z to account for the non-normality in Y– an assumption that is not met here (Kelava et al. 2011). Nevertheless, the results show that Method 2.4 (Latent product) is fairly robust to the levels of non-normality investigated here.

In sum, Chapter 3 found that the preferred moderation methods under normality are the Factor scores and Latent product Methods. The follow-up analyses here demonstrate that these results are robust to levels of non-normality obtained from Chapters 2 and 4. The latent product method performs best in terms of bias and statistical power under conditions of normality and non-normality of X and Z. Future research can investigate even more severe levels of non-normality in X and Z, or can study the impact of non-normal measurement errors of the indicators. Another form of non-normality not investigated here is measurement

with categorical indicators (e.g., Likert scales). The simulations reported here and in Chapter 3 assumed continuous indicators, and future research can investigate the impact of the number of scale points or categories in the indicators.

5.3.3 Study 2c – U-shapes.

Chapter 3 focused on interactions of two, potentially correlated, constructs X and Z but did not investigate U-shaped effects. U- and inverted U-shaped effects manifest when Y increases or decreases when X increases until a minimum or maximum is reached, after which X further decreases or increases (Haans et al. 2016). For instance, Homburg et al. (2011) find among 56 sales managers, 195 sales representatives, and 538 customers that customer orientation (5 indicators, reliability = .88) had an inverted U-shaped relationship with sales performance (3 indicators, reliability = .88).

A common method to test U-shaped relationships is to add a squared term as well as the lower order term to the structural model (Cohen et al. 2003; Haans et al. 2016). Of course, a squared term is an interaction of a variable with itself. Yet the question remains to what extent effects of squared terms have different levels of bias and power than interactions with separate components have (e.g., XZ, hereinafter referred to as interactions).

Reliability of squared terms and standard errors of their effects.

A squared term is an interaction of a variable with itself. One might therefore expect that the reliability of a squared term is higher than that of an interaction. Yet, not only the true scores, but also the measurement errors correlate perfectly (Dimitruk et al. 2007). If X and Z have equal reliabilities, the reliability of a squared term (here X: $r_{XX,XX}$) is usually lower than the reliability of an interaction (cf. Equation 3.4):

$$r_{XX,XX} = r_{X,X}^2, \tag{5.2}$$

where $r_{X,X}$ is the reliability of X (Dimitruk et al. 2007; Moosbrugger et al. 2009). The reliability of a square is equal to the reliability of an interaction when X and Z are

uncorrelated. Yet, the variance of the effect of a squared term is expected to be lower than that of an interaction. The variance of a β regression weight of a focal predictor in a linear model is (Cohen et al. 2003, p. 86):

$$\text{var}(\beta) = \frac{\sigma_Y^2}{\sigma^2} \times \frac{1 - R_{Y.}^2}{n - k - 1} \times \frac{1}{1 - R^2}, \quad (5.3)$$

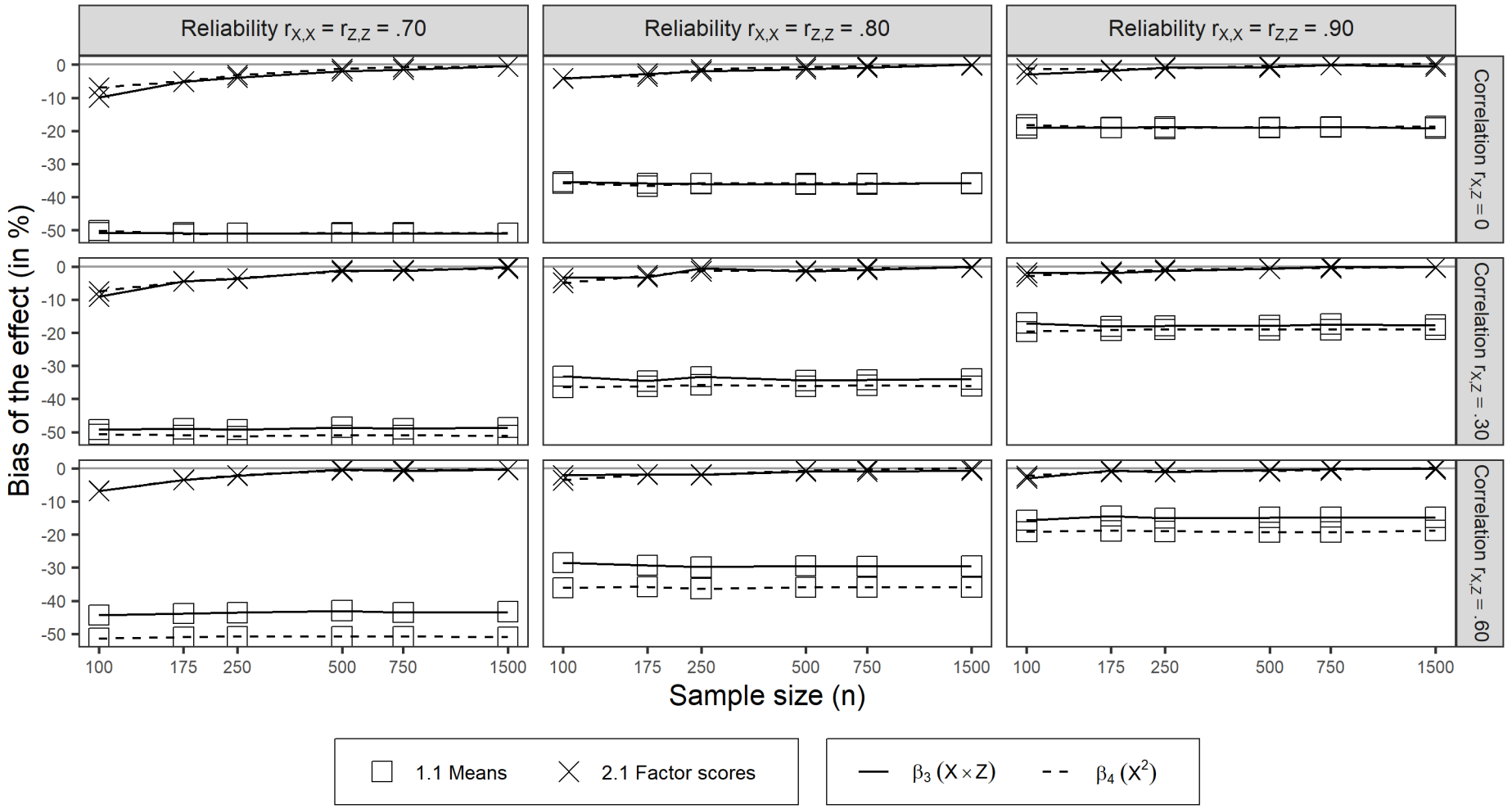
where σ_Y^2 and σ^2 are the respective variances of Y and the focal predictor, $R_{Y.}$ is the multiple correlation of the dependent variable Y with respect to the predictors M and R is the multiple correlation of the focal predictor with any other predictors, n is the sample size, and k is the number of predictors. The product of two standardized normally distributed variables X and Z has mean $r_{X,Z}$, where $r_{X,Z}$ is the correlation between X and Z, and a variance of $1 + r_{X,Z}$. The mean of a square of a standard normal variable X is 1, and the variance 2. Following Equation 5.3, the higher variance of a squared term then leads to a lower variance of its estimate, which increases power. Of course, when X and Z are perfectly correlated, estimating an interaction becomes equal to estimating a squared term as it pertains to the standard error.

Thus, on the one hand, the reliability of squared terms is usually lower than that of interactions, which increases bias and would lead to lower power compared to interactions. On the other hand, the variance of a squared term is higher than that of an interaction, which lowers the variance of the estimate and increases power. Monte Carlo simulations investigated the net differences in power for the effects of squared terms and interactions empirically.

Monte Carlo simulations.

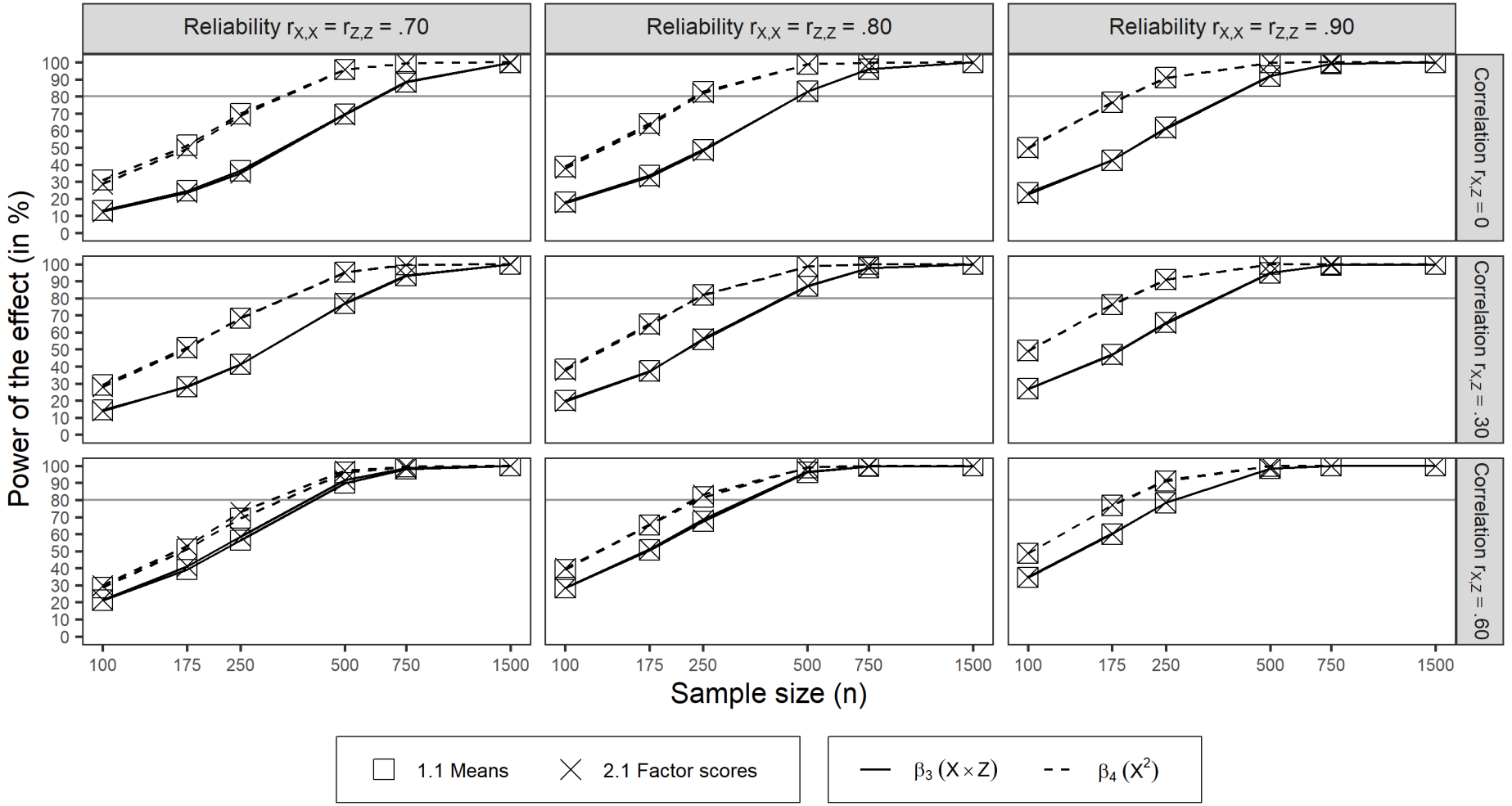
Monte Carlo simulations compared bias and power between squared terms and interactions. Similar to the simulations in Chapter 3, the design with 5,000 replications for each cell varied sample sizes from 100 to 1,500, reliabilities of X and Z with three indicators were .70, .80

Figure 5.6
Monte Carlo: Bias of Interactions and U-Shapes



Notes: Results are based on a Monte Carlo simulation with 5,000 replications. Panels visualize the estimated bias (in %) of the moderation and U-shaped effects across sample sizes (scale is log-transformed) for different reliabilities and correlations between X and Z. Solid horizontal lines indicate a bias of 0%.

Figure 5.7
Monte Carlo: Power of Interactions and U-Shapes



Notes: Results are based on a Monte Carlo simulation with 5,000 replications. Panels visualize the estimated power (in %) of the moderation and U-shaped effects across sample sizes (scale is log-transformed) for different reliabilities and correlations between X and Z. Solid horizontal lines indicate a power of 80%.

and .90, and multicollinearity was fixed to 0, .30 and .60. There were two population models, one for the interaction ($Y_1 = \beta_3 XZ$) and one for the squared term ($Y_2 = \beta_4 X^2$). The effect sizes were held constant: $\beta_3 = \beta_4 = .20$. Method 1.1 (Means), as a benchmark method, and Method 2.1 (Factor scores), as the preferred method in Chapter 3, estimated the U-shaped and interaction effects. The analysis models were $Y_1 = \beta_1 X + \beta_2 Z + \beta_3 XZ$ and $Y_2 = \beta_1 X + \beta_4 X^2$ respectively for the interaction model and the U-shaped model.

Figure 5.6 plots the bias of the estimated effects. First, consistent with Chapter 3, the results show that Method 1.1 (Means) is biased because it inadequately accounts for measurement error in the means of the indicators. Method 2.1 (Factor scores) is able to recover the parameter with minimal bias across sample sizes. The bias increases when reliability of X and Z decreases. Second, as expected, when there is no correlation between X and Z (top row in Figure 5.6), the bias of a squared term is equal to that of an interaction. When the correlation between X and Z increases, the bias of an interaction decreases, due to a higher reliability of the interaction term.

Figure 5.7 plots the estimated statistical power. It shows that the power of a squared term is consistently larger than that of an interaction of the same size. Although the difference becomes smaller with increased correlation between X and Z, it can be substantial. For instance, when the reliability of X and Z is .80, and the correlation between X and Z is .30, the power is about 82% for the squared term, while it is 56% for the interaction term.

In sum, although U-shaped effects might be more biased when measurement error is unaccounted for, their power is higher than that of interaction terms. These simulations assumed however that the structural model is correctly specified. For instance, follow-up studies could investigate the impact of specifying an interaction while the true model has a quadratic term (Ganzach 1997), as well as the effects of multicollinearity and measurement error. Moreover, future research can generalize to moderated U-shapes (Haans et al. 2016).

5.4 Follow-Up Study 3: Discriminant Validity – Multicollinearity

5.4.1 The impact of multicollinearity.

Chapter 4 focused on discriminant validity. Not establishing discriminant validity casts doubt on the validity of process analyses. Nevertheless, finding evidence for discriminant validity does not mean that the analyses are unbiased and make correct inferences. Discriminant validity is commonly a discrete criterion that assesses whether (yes/no) constructs are distinct. Evidence for discriminant validity means support for the construct distinctiveness hypothesis. Lack of evidence for discriminant validity implies insufficient empirical support for construct distinctiveness in a particular study. However, even if measures of constructs meet discrete discriminant validity criteria, small to moderate correlations between them can still be problematic due to multicollinearity.

Multicollinearity, correlations between predictors within a stage in a process model, decreases statistical power. To illustrate this, the variance of a β regression weight of a predictor M in a linear model with standardized variables is (Cohen et al. 2003, p. 86):

$$\sigma^2_{\beta_M} = \frac{1 - R^2_{Y.}}{n - k - 1} \times \text{VIF}, \tag{5.4}$$

$$\text{VIF} = \frac{1}{1 - R^2_{M.M}}, \tag{5.5}$$

where $R_{Y.}$ is the multiple correlation of the dependent variable Y with respect to the predictors M and $R_{M.M}$ is the multiple correlation of the predictors with each other (within or between stages), n is the sample size, and k is the number of predictors. The second term in Equation 5.4 is the variance inflation factor (VIF), shown in Equation 5.5: It is one when there is no multicollinearity and it increases with an increasing multiple correlation of M. Multicollinearity therefore directly increases the variance of β and decreases the statistical power of the parameter significance tests.

5.4.2 Monte Carlo simulations.

A Monte Carlo simulation study with 10,000 replications per cell illustrates this. The population model was $Y = \beta_1 M_1 + \beta_2 M_2 + \varepsilon_Y$, with $\beta_1 = \beta_2 = .20$. The sample size was n = 250 and all variables were standardized and assumed to be measured without error. The correlation between M_1 and M_2 ($r_{M1,M2}$) varied from 0 (no multicollinearity) to .90 in steps of .10. The residual variance var(ε_Y) was set such that the multiple correlation of Y was kept constant ($R^2_{Y.} = .08$). The estimation model was a linear regression of Y on M_1 and M_2.

Table 5.5 has the results. The horizontal dashed line distinguishes absence of multicollinearity (above the dashed line) from the presence of multicollinearity (below the dashed line). Columns B and C in Table 5.5 show that the population effects of .20 were estimated accurately, regardless of the level of multicollinearity (Column A), but that the standard errors increase when multicollinearity increases (columns E and F). For example, increasing the correlation between M_1 and M_2 from 0 to .50, a moderate correlation, increases the average standard error about 40%, from .061 to .086. This also decreases statistical power at the current sample size of n = 250 from 90% to 64% (columns F and G). Correlations higher than .30 yielded a power smaller than 80%. Even higher correlations further lower statistical power to an estimated 18% for a multicollinearity level of .90. This even occurs when the VIF (reported between parentheses in Column A) is below conventional rules of

Table 5.5
Multicollinearity Decreases Statistical Power

(A) $r_{X1,X2}$ (VIF)	(B) $\hat{\beta}_1$	(C) $\hat{\beta}_2$	(D) $\widehat{SE}(\hat{\beta}_1)$	(E) $\widehat{SE}(\hat{\beta}_2)$	(F) $\widehat{Power}(\hat{\beta}_1)$	(G) $\widehat{Power}(\hat{\beta}_2)$
0 (1)	.200	.200	.061	.061	90%	90%
.10 (1.01)	.201	.200	.064	.064	88%	87%
.20 (1.04)	.201	.199	.068	.068	84%	83%
.30 (1.10)	.201	.200	.073	.073	79%	78%
.40 (1.19)	.200	.201	.079	.079	72%	72%
.50 (1.33)	.200	.202	.086	.086	64%	65%
.60 (1.56)	.199	.201	.096	.097	54%	55%
.70 (1.96)	.200	.200	.111	.111	43%	43%
.80 (2.78)	.199	.200	.136	.137	31%	31%
.90 (5.26)	.200	.198	.193	.193	18%	18%

Notes: Table contains design and results from a Monte Carlo analysis with n = 250 and 10,000 replications per cell. Columns B to E are averages across the replications, and power (columns F and G) was the proportion of replications that had parameter estimate with $p < .05$.

thumb such as 4 and 10, that correspond to correlations of .87 and .95 respectively. These results show that although measures of constructs in a process model may express distinctiveness, correlations between predictors can still lead to severe drops in statistical power. The effect of multicollinearity on the power is continuous compared to the often discrete discriminant validity criteria. Low power can even occur at low to moderate levels of correlations between predictors.[13]

5.4.3 Discussion.

In sum, this dissertation discussed three consequences of multicollinearity. First, and surprisingly, Chapter 3 showed that multicollinearity between lower order terms increases the reliability of a product term, which decreases the attenuation bias in moderation effects due to not accounting for measurement error. Moreover, it increases the statistical power of finding a true moderation effect. Second, Chapter 4 conceptualized multicollinearity as correlated measures of constructs on the right side of an equation in a process model. Of course, high levels of multicollinearity between measures of constructs are at risk of not meeting bivariate and multivariate discriminant validity criteria. Finally, the results of the follow-up analysis in Table 5.5 revealed that low to moderate levels of multicollinearity decreased statistical power in multiple regression but did not lead to estimation bias.

After following up on several issues, the remainder of this dissertation zooms out to discuss the breadth of process analysis. It concludes with potential next steps for usage of process analysis in marketing research.

5.5 The Breadth of Process Theories

Process analysis aims to quantify the pathways through which inputs have effects on output variables. It is a general statistical analysis method to investigate relationships between any

[13] Severe multicollinearity was also found to result in model non-convergence, biased estimates, and Type I error. Mason and Perreault (1991), Grewal et al. (2004), and Kalnins (2018) further explore the consequences of (more severe) multicollinearity.

number of measures of constructs, as well as mediation and moderation hypotheses. Process analysis is a useful tool to test broad marketing theories. For example, one of the multiple mediation studies that was investigated in Chapter 4 of this dissertation estimated the effect of choice difficulty on choice delegation through four mediators: rated unattractiveness, rated difficulty, anticipated disappointment and anticipated regret (Steffel and Williams 2018). As another example, Völckner and Sattler (2006) investigated the effects of ten determinants of brand extension success and hypothesized ten mediating and five moderating relationships. In comparison, early applications of process analysis by Wright (1921) had a total of ten constructs and six structural equations to determine the weight of guinea pigs at birth.

However, the question remains how broad contemporary marketing theories are, and to what extent their breadth has grown over time. Table 5.6 summarizes existing data on theory breadth in marketing and business research to investigate this. It measures breadth by the number of constructs in theories. Hulland et al. (1996) present an early review of 186 structural equation models (SEMs) published in 11 marketing journals between 1980 and 1994. It found a mean of 6.9 constructs, which was decomposed in 2.7 inputs, 2.5 mediators, and 1.7 outcomes. The average process model in that sample investigated one or two outcomes of interest but specified the effects of multiple inputs through multiple mediators, akin to the contemporary multiple mediation models studied in Chapter 4. Interestingly, Hulland et al. (1996) did not find a difference in the number of constructs included in the models between 1980-1989 and 1990-1994. Similarly, Baumgartner and Homburg (1996) reviewed SEMs in four marketing journals between 1977 and 1994. They found a median number of latent constructs of 5, which is substantively smaller than the mean of 6.9 in Hulland et al. (1996). Yet, Martínez-López et al. (2013) directly followed up on Baumgartner and Homburg (1996) by investigating 191 articles in the same four marketing journals from

1995 to 2007. The median number of latent constructs in articles published between 1995 and 2007 was 7, which is larger than the median of 5 found for 1977-1994 articles.

Hair et al. (2012a) investigated the use of partial least squares (PLS), an estimation algorithm for process analysis, in 204 articles published in 24 marketing journals between 1981 and 2010. The review found a mean of 7.9 and a median of 7 constructs, a similar result to that from Martínez-López et al. (2013). Importantly, the mean was 6.3 and the median was 6 for articles published before 2000, whereas for articles published after 2000 the mean number of constructs was 8.4 and the median was 8 ($p < .01$ for the difference in means before and after 2000). Similar results were found in the strategic management research domain (Hair et al. 2012b). Another review of PLS models studied 191 articles that were published in *Management Information Systems Quarterly (MISQ)* between 2012 and 2015 (Hair et al. 2017). It found a mean of 8.8 and median of 8 constructs which is slightly higher than the mean of 8.1 and median of 7 found in a review of the same outlet between 1992-2001 by Ringle et al. (2012). Although these findings indicate a positive trend in theory breadth over time, a review of 37 SEMs in eight accounting journals by Herda (2013) concluded the opposite. A follow-up analysis of the data in the Appendix of Herda (2013) found a negative correlation between the natural logarithm of the number of constructs in a theory with publication year ($r = -.32$, $p = .06$).

Follow-up analyses of the 2000-2017 data on moderation tests in the *Journal of Marketing (JM)* and *Journal of Marketing Research (JMR)* in Chapter 3 of this dissertation revealed a non-significant negative correlation between the natural logarithm of the number of moderation tests per article and the publication year ($r = -.15$, $p = .15$; mean of 5.2 and median of 4 moderation tests per article). However, there was a strong positive correlation between the natural logarithm of the total number of predictors in the latent moderation equation and the year of publication ($r = .50$, $p < .01$).

Table 5.6
Selected Data on Theory Breadth in Business and Marketing Research

Article	Time period investigated	Number of articles and outlets	The number (#) of constructs in the theory	The number (#) of constructs in the theory over time
Hulland et al. (1996)	1980-1994	186 articles in 11 marketing journals including *JM*, *JMR*, *JCR*, *IJRM*	Mean = 6.9 (2.7 X, 2.5 M, 1.7 Y)	Non-significant difference between 1980-1989 and 1990-1994
Baumgartner and Homburg (1996)	1977-1994	149 articles in 4 marketing journals: *JM*, *JMR*, *JCR*, *IJRM*	Median = 5	Compare with Martínez-López et al. (2013)
Boyd et al. (2005)	1998-2000	196 articles in 4 management journals	# of X per analysis: Mean = 5.7, Median = 4	NA
Shah and Goldstein (2006)	1984-2003	93 articles in 4 operations management journals	Mean = 4.4, Median = 4	NA
Martínez-López et al. (2013)	1995-2007	191 articles in 4 marketing journals: *JM*, *JMR*, *JCR*, *IJRM*	Median = 7	Compare with Baumgartner and Homburg (1996)
Hair et al. (2012a)	1981-2010	204 articles in 24 marketing journals including *JM*, *JMR*, *JCR*, *IJRM*	Mean = 7.9, Median = 7	< 2000: Mean = 6.3, Median = 6 ≥ 2000: Mean = 8.4, Median = 8
Hair et al. (2012b)	1981-2010	37 articles in 8 management journals	Mean = 7.5, Median = 6	< 2000: Mean = 7, Median = 6 ≥ 2000: Mean = 8.1, Median = 6
Ringle et al. (2012)	1992-2011	65 articles in 1 information systems journal: *MISQ*	Mean = 8.1, Median = 7	Compare with Hair et al. (2017)
Herda (2013)	2000-2011	37 articles in 8 accounting journals	Mean = 5.4, Median = 5	r of ln(# of constructs) with publication year = -.32 (p = .06)
Hair et al. (2017)	2012-2015	191 articles in 1 information systems journal: *MISQ*	Mean = 8.8, Median = 8	Compare with Ringle et al. (2012)
Pieters (2017)	2014-2016	138 articles in 1 marketing journal: *JCR*	Out of 166 mediation analyses: 1 M: 55 (33%) >1 M: 29 (17%)	NA
Chapter 3 of this dissertation	2000-2017	97 articles in 2 marketing journals: *JM*, *JMR*	# of moderation tests per article: Mean = 5.2, Median = 4	r of ln(# of moderation tests per article) with publication year = -.15 (p = .15)
			# of predictors: Mean = 14.8, Median = 13	r of ln(# of predictors) with publication year = .50 (p < .01)
Chapter 4 of this dissertation	2017-2019	23 studies in 15 articles in 3 marketing journals: *JM*, *JMR*, *JCR*	# of X per study: Mean = 1.04, Median = 1	NA
			# of M per study: Mean = 2.4, Median = 2	
			# of Y per study Mean = 1.2, Median = 1	

Notes: NA means that the information was unavailable in the meta-analysis report. X refers to input, M to mediator and Y to output, and r refers to a correlation. *JM* is the *Journal of Marketing*, *JMR* is the *Journal of Marketing Research*, *JCR* is the *Journal of Consumer Research*, *IJRM* is the *International Journal of Research in Marketing* and *MISQ* is *Management Information Systems Quarterly*.

Although increasing theory breadth may in part reflect an increasing amount of control variables in moderation models, an upward trend in control variables is unlikely to fully account for the increasing theory breadth. Boyd et al. (2005) found in a review of measurement practices in strategic management that more than 99% of control variables were single-indicators. Generally, the data summarized in Table 5.6 reported the number of latent constructs in the model, reflected by multiple indicators. Hair et al. (2012a) also concluded that the number of models with single-item constructs decreased from 61% in their sample before 2000 to 42% in 2000 and onward. Table 5.6 contains additional details.

In sum, the body of evidence indicates that the breadth of marketing theories has increased over time. Multiple explanations, speculatively, may account for this. First, there might have been a shift towards research questions that warrant broader theorization. There can also be beliefs among researchers that broader theories are more interesting ones. As the field has been progressing, marketing researchers might have contributed by extending existing theories with additional process variables. Yet, looking at novel phenomena through the lens of existing theories "…may lead us to borrow imperfect theories rather than develop fresh ones" which "…may cause researchers to complicate theories as they adapt them to the new context" (Tellis 2017, p. 3). Second, methodological advances in process analysis may have facilitated empirical testing of broad theories. For instance, covariance structure analysis "…is explicitly aimed at complex testing of theory" and "…makes possible the rigorous testing of theories that have until now been very difficult to test adequately" (Kerlinger 1977, p. 9). Accessible implementations, such as in R (Rosseel 2012), further contribute to this.

Third, research with broad and nuanced theories may be increasingly selected into publication, whereas studies with relatively narrow theories remain unpublished. Publication outlets have become more selective over time, which is reflected in decreasing acceptance rates. For *JMR*, the acceptance rate dropped from 15% to 12% between 2006 and 2012

(Huber and Erdem 2014), and further decreased below 10% in 2017 (Grewal 2017). The acceptance rate of *JM* dropped from 10.2% between 1993 to 1996 to 8% in 2017 (Moorman et al. 2019; Varadarajan 1996). As a corollary, this might have increased the demand for conceptual rigor and theoretical contributions (Russell-Bennett and Baron 2015). Such theoretical contributions are unlikely to be accomplished if existing theories are simply applied in new settings. As Whetten (1989) notes: "…theorists need to learn something new about the theory itself as a result of working with it under different *condition*s" (p. 493, emphasis added). That likely requires the addition of new moderators to process models. The final section discusses challenges and opportunities for process models to test increasingly broad marketing theories.

5.6 Testing Broad Theories with Process Analysis: The Road Ahead

The increasing theoretical breadth poses challenges for the road ahead. Broad theories risk failing to establish discriminant validity. Everything else held equal, adding a construct decreases the likelihood that all constructs are theoretically distinct from the others and that their measures have unique variance not accounted for by the other measures. Lack of discriminant validity leads to construct proliferation, the accumulation of ostensibly different but potentially identical constructs (Shaffer et al. 2016). It occurs when theoretically and/or empirically indistinct constructs receive different labels. For instance, items that measured claim believability, message believability and trustworthiness of an ad were also used in ad credibility measures (Bergkvist and Langner 2019). The extent of this item overlap suggests that these constructs might be virtually identical to each other, and that the ad credibility construct was proliferated.

Mediators are of particular risk to be redundant. Serial mediators are by definition hypothesized to be related to other mediators, and perhaps relate to inputs and outputs separately. Multiple mediators in parallel might set out to capture fine-grained processes that

cannot be distinguished theoretically or empirically. In this case, construct proliferation can even lead to *process proliferation*, the accumulation of theoretically or empirically indistinguishable processes. As an example of potential process proliferation in simple mediation, Bove et al. (2009) studied the influence of a customer's commitment to a service worker on customer organizational citizen behaviors (OCBs). An analysis of 484 customers in three service contexts (pharmacy, hairdressing, and medical services) found an effect of commitment on customer OCBs, partially mediated by personal loyalty. However, Farrell (2010) challenged the discriminant validity of these constructs, particularly between commitment and the personal loyalty mediator. Here, lack of discriminant validity results in the possibility of process proliferation. It provides evidence against the hypothesis of the multiple processes that drive the effect of commitment on customer OCBs: the mediation effect through personal loyalty, and the direct non-mediated effect of commitment on customer OCB that circumvents personal loyalty. Chapter 4 of this dissertation presented additional case studies of multiple mediation.

Yet, broad theories also provide opportunities for marketing research. They enable a better and nuanced understanding of complex real-world phenomenon and account for factors that would bias the results or limit generalizability if omitted. To continue moving forward, it is perhaps best to pay particular attention to the development of meaningful theories. Theory building has an inherent tension between parsimony and comprehensiveness (Whetten 1989). Parsimony and simplicity, which are metatheoretical criteria (Gawronski and Bodenhausen 2015) or virtues of theory building (Quine and Ullian 1978), prefer a theory with fewer constructs over one with more constructs, holding everything else equal. Simplicity promotes brevity and interesting and impactful research (Tellis 2017). On the other hand, generality and comprehensiveness prefer theories with more rather than less explanatory breadth (Gawronski and Bodenhausen 2015). Thoughtful theory building has important tradeoffs:

simplicity prevents construct and process proliferation, while broader theories contribute to generality.

From an empirical perspective, two concurrent developments provide opportunities for valid process analyses in the marketing discipline as a whole. First, Chapters 3 and 4 of this dissertation provided evidence that measurement reliability has increased over time. Chapter 3 found mean reliabilities of .86 of multi-item inputs, moderators and outputs in 504 moderation tests in 97 articles published in *JM* and *JMR* between 2000 and 2017. Moreover, Chapter 4 found mean reliabilities of .88 to .91 for inputs, mediators and outcomes in 23 recent multiple mediation studies in *JM*, *JMR*, and *JCR*. These estimates are substantively higher than the mean reliability of .77 found in an early meta-analysis of 4,286 measures in 832 marketing articles published between 1960 and 1992 (Peterson 1994). It might reflect an increased proficiency in scale construction or better selection of existing reliable measures over unreliable ones.

Second, samples have become bigger. For instance, Martínez-López et al. (2013) found a median sample size of 259 between 1995 and 2007, substantively larger than the median of 178 between 1977 and 1994 in the same set of marketing journals (Baumgartner and Homburg 1996). These developments might imply a growth in statistical power of marketing research, like psychology research, which has experienced a slight growth in statistical power over time (Rossi 1990; Szucs and Ioannidis 2017b). As shown in Chapter 4 of this dissertation, reliable measures and large samples contribute to the distinctiveness condition of meaningful process analysis (Pieters 2017), which enables meaningful testing of broad marketing theories. In conclusion, the empirical reliability and discriminant validity criteria of construct validity are more likely attained (Peter 1981). Furthermore, a meta-analysis of 176 marketing meta-analyses found that the average effect size in marketing has been increasing over time (Eisend 2015). It implies good nomological validity.

The year 2020 marks the centennial anniversary of seminal process analysis contributions (Wright 1920). To continue moving forward, thoughtful theory building that trades off parsimony and comprehensiveness can result in meaningful process theories and strong theoretical contributions. Process analysis methodologies have become well-equipped to quantify the pathways in such relevant marketing theories. Hopefully, the essays in this dissertation further contribute to the usefulness and validity of process analysis methods and applications in marketing research.

References

Aguinis, Herman, James C. Beaty, Robert J. Boik, and Charles A. Pierce (2005), "Effect Size and Power in Assessing Moderating Effects of Categorical Variables Using Multiple Regression: A 30-Year Review," *Journal of Applied Psychology*, 90 (1), 94-107.

Ahrens, Jan, James R. Coyle, and Michal Ann Strahilevitz (2013), "Electronic Word of Mouth: The Effects of Incentives on E-Referrals by Senders and Receivers," *European Journal of Marketing*, 47 (7), 1034-51.

Alexandrov, Aliosha, Bryan Lilly, and Emin Babakus (2013), "The Effects of Social- and Self-Motives on the Intentions to Share Positive and Negative Word of Mouth," *Journal of the Academy of Marketing Science*, 41 (5), 531-46.

Algina, James and Bradley C. Moulder (2001), "A Note on Estimating the Jöreskog-Yang Model for Latent Variable Interaction Using LISREL 8.3," *Structural Equation Modeling: A Multidisciplinary Journal*, 8 (1), 40-52.

Anderson, Eugene W. (1998), "Customer Satisfaction and Word of Mouth," *Journal of Service Research*, 1 (1), 5-17.

Anderson, James C. and David W. Gerbing (1988), "Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach," *Psychological Bulletin*, 103 (3), 411-23.

Asparouhov, Tihomir and Bengt Muthén (2020), "Bayesian Estimation of Single and Multilevel Models with Latent Variable Interactions," *forthcoming in Structural Equation Modeling: A Multidisciplinary Journal*, 1-15.

Auh, Seigyoung, Bulent Menguc, Constantine S. Katsikeas, and Yeon Sung Jung (2019), "When Does Customer Participation Matter? An Empirical Investigation of the Role of Customer Empowerment in the Customer Participation–Performance Link," *Journal of Marketing Research*, 56 (6), 1012-33.

Bagozzi, Richard P. and Youjae Yi (1988), "On the Evaluation of Structural Equation Models," *Journal of the Academy of Marketing Science*, 16 (1), 74-94.

Baker, Wayne E. and Nathaniel Bulkley (2014), "Paying It Forward vs. Rewarding Reputation: Mechanisms of Generalized Reciprocity," *Organization Science*, 25 (5), 1493-510.

Baron, Reuben M. and David A. Kenny (1986), "The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology*, 51 (6), 1173-82.

Baumgartner, Hans and Christian Homburg (1996), "Applications of Structural Equation Modeling in Marketing and Consumer Research: A Review," *International Journal of Research in Marketing*, 13 (2), 139-61.

Baumgartner, Hans and Bert Weijters (2017), "Measurement Models for Marketing Constructs," in *Handbook of Marketing Decision Models*, Berend Wierenga and Ralf van der Lans, eds. 2nd ed. Cham, Switzerland: Springer.

Bellezza, Silvia, Neeru Paharia, and Anat Keinan (2017), "Conspicuous Consumption of Time: When Busyness and Lack of Leisure Time Become a Status Symbol," *Journal of Consumer Research*, 44 (1), 118-38.

Berger, Jonah (2014), "Word of Mouth and Interpersonal Communication: A Review and Directions for Future Research," *Journal of Consumer Psychology*, 24 (4), 586-607.

Bergkvist, Lars and Tobias Langner (2019), "Construct Heterogeneity and Proliferation in Advertising Research," *International Journal of Advertising*, 38 (8), 1286-302.

Bergkvist, Lars and John R. Rossiter (2007), "The Predictive Validity of Multiple-Item Versus Single-Item Measures of the Same Constructs," *Journal of Marketing Research*, 44 (2), 175-84.

Blalock, H. M. (1965), "Some Implications of Random Measurement Error for Causal Inferences," *American Journal of Sociology*, 71 (1), 37-47.

Bollen, Kenneth A. (1989), Structural Equations with Latent Variables. New York: Wiley.

Bougie, Roger, Rik Pieters, and Marcel Zeelenberg (2003), "Angry Customers Don't Come Back, They Get Back: The Experience and Behavioral Implications of Anger and Dissatisfaction in Services," *Journal of the Academy of Marketing Science*, 31 (4), 377-93.

Bove, Liliana L., Simon J. Pervan, Sharon E. Beatty, and Edward Shiu (2009), "Service Worker Role in Encouraging Customer Organizational Citizenship Behaviors," *Journal of Business Research*, 62 (7), 698-705.

Boyd, Brian K., Steve Gove, and Michael A. Hitt (2005), "Construct Measurement in Strategic Management Research: Illusion or Reality?," *Strategic Management Journal*, 26 (3), 239-57.

Brandt, Holger, Augustin Kelava, and Andreas Klein (2014), "A Simulation Study Comparing Recent Approaches for the Estimation of Nonlinear Effects in SEM under the Condition of Nonnormality," *Structural Equation Modeling: A Multidisciplinary Journal*, 21 (2), 181-95.

Bronner, Fred and Robert De Hoog (2011), "Vacationers and eWOM: Who Posts, and Why, Where, and What?," *Journal of Travel Research*, 50 (1), 15-26.

Brown, Tom J., Thomas E. Barry, Peter A. Dacin, and Richard F. Gunst (2005), "Spreading the Word: Investigating Antecedents of Consumers' Positive Word-of-Mouth Intentions and Behaviors in a Retailing Context," *Journal of the Academy of Marketing Science*, 33 (2), 123-38.

Burks, Barbara Stoddard (1928), "The Relative Influence of Nature and Nurture Upon Mental Development; a Comparative Study of Foster Parent-Foster Child Resemblance and True

Parent-True Child Resemblance," in *The Twenty-Seventh Yearbook of the National Society for the Study of Education*, Guy Montrose Whipple, ed. Bloomington, IL: Public School Publishing Company.

Burt, Cyril (1943), "Validating Tests for Personnel Selection," *British Journal of Psychology, General Section*, 34 (1), 1-19.

Busemeyer, Jerome R. and Lawrence E. Jones (1983), "Analysis of Multiplicative Combination Rules When the Causal Variables Are Measured with Error," *Psychological Bulletin*, 93 (3), 549-62.

Campbell, Donald T. and Donald W. Fiske (1959), "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," *Psychological Bulletin*, 56 (2), 81-105.

Campbell, Margaret C. and Karen Page Winterich (2018), "A Framework for the Consumer Psychology of Morality in the Marketplace," *Journal of Consumer Psychology*, 28 (2), 167-79.

Chang, Winston, Joe Cheng, J. J. Allaire, Yihui Xie, and Jonathan McPherson (2019), "Shiny: Web Application Framework for R."

Charter, Richard A. and Barbara S. Larsen (1983), "Fisher's Z to r," *Educational and Psychological Measurement*, 43 (1), 41-42.

Chen, Zoey and Jonah Berger (2016), "How Content Acquisition Method Affects Word of Mouth," *Journal of Consumer Research*, 43 (1), 86-102.

Cohen, Jacob (1988), Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, Jacob, Patricia Cohen, Stephen G. West, and Leona S. Aiken (2003), Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, Maxime C., Carlos Fernández, and Anindya Ghose (2019), "Empirical Analysis of Referrals in Ride-Sharing," [available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3345669].

Cole, David A. and Kristopher J. Preacher (2014), "Manifest Variable Path Analysis: Potentially Serious and Misleading Consequences Due to Uncorrected Measurement Error," *Psychological Methods*, 19 (2), 300-15.

Cortina, Jose M., Hannah M. Markell-Goldstein, Jennifer P. Green, and Yingyi Chang (2019), "How Are We Testing Interactions in Latent Variable Models? Surging Forward or Fighting Shy?," *forthcoming in Organizational Research Methods*, 1-29.

Court, Andrew T. (1930), "Measuring Joint Causation," *Journal of the American Statistical Association*, 25 (171), 245-54.

Cronbach, Lee J. and Paul E. Meehl (1955), "Construct Validity in Psychological Tests," *Psychological Bulletin*, 52 (4), 281-302.

Crow, James F. (1992), "Sewall Wright's Place in Twentieth-Century Biology," in *The Founders of Evolutionary Genetics: A Centenary Reappraisal*, Sahotra Sarkar, ed. Dordrecht, The Netherlands: Springer.

Curran, Patrick J., Stephen G. West, and John F. Finch (1996), "The Robustness of Test Statistics to Nonnormality and Specification Error in Confirmatory Factor Analysis," *Psychological Methods*, 1 (1), 16-29.

Dahlquist, Steven H. and David A. Griffith (2014), "Multidyadic Industrial Channels: Understanding Component Supplier Profits and Original Equipment Manufacturer Behavior," *Journal of Marketing*, 78 (4), 59-79.

De Luca, Luigi M. and Kwaku Atuahene-Gima (2007), "Market Knowledge Dimensions and Cross-Functional Collaboration: Examining the Different Routes to Product Innovation Performance," *Journal of Marketing*, 71 (1), 95-112.

De Matos, Celso A. and Carlos A. V. Rossi (2008), "Word-of-Mouth Communications in Marketing: A Meta-Analytic Review of the Antecedents and Moderators," *Journal of the Academy of Marketing Science*, 36 (4), 578-96.

Deighton, John, Debbie MacInnis, Ann McGill, and Baba Shiv (2010), "Broadening the Scope of Consumer Research," *Journal of Consumer Research*, 36 (6), v-vii.

Devlieger, Ines, Axel Mayer, and Yves Rosseel (2016), "Hypothesis Testing Using Factor Score Regression: A Comparison of Four Methods," *Educational and Psychological Measurement*, 76 (5), 741-70.

Dimitruk, Polina, Karin Schermelleh-Engel, Augustin Kelava, and Helfried Moosbrugger (2007), "Challenges in Nonlinear Structural Equation Modeling," *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 3 (3), 100-14.

Downing, Charles E. (1999), "System Usage Behavior as a Proxy for User Satisfaction: An Empirical Investigation," *Information & Management*, 35 (4), 203-16.

Duncan, Otis Dudley (1966), "Path Analysis: Sociological Examples," *American Journal of Sociology*, 72 (1), 1-16.

Eggert, Andreas, Lena Steinhoff, and Carina Witte (2019), "Gift Purchases as Catalysts for Strengthening Customer–Brand Relationships," *Journal of Marketing*, 83 (5), 115-32.

Eisend, Martin (2015), "Have We Progressed Marketing Knowledge? A Meta-Meta-Analysis of Effect Sizes in Marketing Research," *Journal of Marketing*, 79 (3), 23-40.

Engel, James F., Robert J. Kegerreis, and Roger D. Blackwell (1969), "Word-of-Mouth Communication by the Innovator," *Journal of Marketing*, 33 (3), 15-19.

Farrell, Andrew M. (2010), "Insufficient Discriminant Validity: A Comment on Bove, Pervan, Beatty, and Shiu (2009)," *Journal of Business Research*, 63 (3), 324-27.

Finch, John F., Stephen G. West, and David P. MacKinnon (1997), "Effects of Sample Size and Nonnormality on the Estimation of Mediated Effects in Latent Variable Models," *Structural Equation Modeling: A Multidisciplinary Journal*, 4 (2), 87-107.

Finney, Sara J. and Christine DiStefano (2006), "Non-Normal and Cateogrical Data in Structural Equation Modeling," in *Structural Equation Modeling: A Second Course*, Gregory R. Hancock and Ralph O. Mueller, eds. Greenwich, Conneticut: IAP.

Flake, Jessica K., Jolynn Pek, and Eric Hehman (2017), "Construct Validation in Social and Personality Research: Current Practice and Recommendations," *Social Psychological and Personality Science*, 8 (4), 370-78.

Flynn, Leisa R., Ronald E. Goldsmith, and Jacqueline K. Eastman (1996), "Opinion Leaders and Opinion Seekers: Two New Measurement Scales," *Journal of the Academy of Marketing Science*, 24 (2), 137-47.

Foldnes, Njål and Knut Arne Hagtvet (2014), "The Choice of Product Indicators in Latent Variable Interaction Models: Post Hoc Analyses," *Psychological Methods*, 19 (3), 444-57.

Fornell, Claes and David F. Larcker (1981), "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research*, 18 (1), 39-50.

Franke, George and Marko Sarstedt (2018), "Heuristics Versus Statistics in Discriminant Validity Testing: A Comparison of Four Procedures," *Internet Research*, 29 (3), 430-47.

Frick, Robert W. (1996), "The Appropriate Use of Null Hypothesis Testing," 1 (4), 379-90.

Friedman, Lynn and Melanie Wall (2005), "Graphical Views of Suppression and Multicollinearity in Multiple Linear Regression," *The American Statistician*, 59 (2), 127-36.

Friestad, Marian and Peter Wright (1995), "Persuasion Knowledge: Lay People's and Researchers' Beliefs About the Psychology of Advertising," *Journal of Consumer Research*, 22 (1), 62-74.

Fuller, Wayne A. and Michael A. Hidiroglou (1978), "Regression Estimation after Correcting for Attenuation," *Journal of the American Statistical Association*, 73 (361), 99-104.

Fürst, Andreas, Martin Leimbach, and Jana-Kristin Prigge (2017), "Organizational Multichannel Differentiation: An Analysis of Its Impact on Channel Relationships and Company Sales Success," *Journal of Marketing*, 81 (1), 59-82.

Ganzach, Yoav (1997), "Misleading Interaction and Curvilinear Terms," *Psychological Methods*, 2 (3), 235-47.

Gawronski, Bertram and Galen V. Bodenhausen (2015), "Theory Evaluation," in *Theory and Explanation in Social Psychology*, Bertram Gawronski and Galen V. Bodenhausen, eds. New York: Guilford.

Gaylord, R. H. and J. B. Carroll (1948), "A General Approach to the Problem of the Population Control Variable," *American Psychologist*, 3 (310), 209-22.

Gelman, Andrew and Eric Loken (2014), "The Statistical Crisis in Science: Data-Dependent Analysis—a "Garden of Forking Paths"—Explains Why Many Statistically Significant Comparisons Don't Hold Up," *American Scientist*, 102 (6), 460-66.

Goel, Sharad, Ashton Anderson, Jake Hofman, and Duncan J. Watts (2015), "The Structural Virality of Online Diffusion," *Management Science*, 62 (1), 180-96.

Goenka, Shreyans and Stijn M. J. Van Osselaer (2019), "Charities Can Increase the Effectiveness of Donation Appeals by Using a Morally Congruent Positive Emotion," *Journal of Consumer Research*, 46 (4), 774-90.

Goldsby, Thomas J., A. Michael Knemeyer, Jason W. Miller, and Carl Marcus Wallenburg (2013), "Measurement and Moderation: Finding the Boundary Conditions in Logistics and Supply Chain Research," *Journal of Business Logistics*, 34 (2), 109-16.

Greene, William H. (2008), Econometric Analysis (6th ed.). Upper Saddle River, NJ: Pearson.

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman (2016), "Statistical Tests, *P* Values, Confidence Intervals, and Power: A Guide to Misinterpretations," *European Journal of Epidemiology*, 31 (4), 337-50.

Grewal, Lauren and Andrew T. Stephen (2019), "In Mobile We Trust: The Effects of Mobile Versus Nonmobile Reviews on Consumer Purchase Intentions," *Journal of Marketing Research*, 56 (5), 791-808.

Grewal, Rajdeep (2017), "*Journal of Marketing Research*: Looking Forward," *Journal of Marketing Research*, 54 (1), 1-4.

Grewal, Rajdeep, Joseph A. Cote, and Hans Baumgartner (2004), "Multicollinearity and Measurement Error in Structural Equation Models: Implications for Theory Testing," *Marketing Science*, 23 (4), 519-29.

Grewal, Rajdeep, Alok Kumar, Girish Mallapragada, and Amit Saini (2013), "Marketing Channels in Foreign Markets: Control Mechanisms and the Moderating Role of Multinational Corporation Headquarters–Subsidiary Relationship," *Journal of Marketing Research*, 50 (3), 378-98.

Grice, James W. (2001), "Computing and Evaluating Factor Scores," *Psychological Methods*, 6 (4), 430-50.

Haans, Richard F. J., Constant Pieters, and Zi-Lin He (2016), "Thinking About U: Theorizing and Testing U- and Inverted U-Shaped Relationships in Strategy Research," *Strategic Management Journal*, 37 (7), 1177-95.

Hair, Joe F., Marko Sarstedt, Christian M. Ringle, and Jeannette A. Mena (2012a), "An Assessment of the Use of Partial Least Squares Structural Equation Modeling in Marketing Research," *Journal of the Academy of Marketing Science*, 40 (3), 414-33.

Hair, Joe, Carole L. Hollingsworth, Adriane B. Randolph, and Alain Yee Loong Chong (2017), "An Updated and Expanded Assessment of PLS-SEM in Information Systems Research," *Industrial Management & Data Systems*, 117 (3), 442-58.

Hair, Joseph F., Marko Sarstedt, Torsten M. Pieper, and Christian M. Ringle (2012b), "The Use of Partial Least Squares Structural Equation Modeling in Strategic Management Research: A Review of Past Practices and Recommendations for Future Applications," *Long Range Planning*, 45 (5-6), 320-40.

Harter, James K. and Frank L. Schmidt (2008), "Conceptual Versus Empirical Distinctions among Constructs: Implications for Discriminant Validity," *Industrial and Organizational Psychology*, 1 (1), 36-39.

Hartline, Michael D. and Keith C. Jones (1996), "Employee Performance Cues in a Hotel Service Environment: Influence on Perceived Service Quality, Value, and Word-of-Mouth Intentions," *Journal of Business Research*, 35 (3), 207-15.

Haws, Kelly L., Utpal M. Dholakia, and William O. Bearden (2010), "An Assessment of Chronic Regulatory Focus Measures," *Journal of Marketing Research*, 47 (5), 967-82.

Hayes, Andrew F. and Kristopher J. Preacher (2013), "Conditional Process Modeling: Using Structural Equation Modeling to Examine Contingent Causal Processes," in *Structural Equation Modeling: A Second Course*, Gregory R. Hancock and Ralph O. Mueller, eds. 2nd ed. Greenwich, CT: Information Age Publishing.

Held, Leonhard and Manuela Ott (2018), "On *P*-Values and Bayes Factors," *Annual Review of Statistics and Its Application*, 5 (1), 393-419.

Henseler, Jörg, Christian M. Ringle, and Marko Sarstedt (2015), "A New Criterion for Assessing Discriminant Validity in Variance-Based Structural Equation Modeling," *Journal of the Academy of Marketing Science*, 43 (1), 115-35.

Herda, David N. (2013), "Structural Equation Modeling in the Accounting Literature: Observations and Suggestions," *Journal of Theoretical Accounting Research*, 8 (2), 103-38.

Hoch, Stephen J. and Young-Won Ha (1986), "Consumer Learning: Advertising and the Ambiguity of Product Experience," *Journal of Consumer Research*, 13 (2), 221-33.

Hoenig, John M. and Dennis M. Heisey (2001), "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis," *The American Statistician*, 55 (1), 19-24.

Homburg, Christian and Matthias Bucerius (2005), "A Marketing Perspective on Mergers and Acquisitions: How Marketing Integration Affects Postmerger Performance," *Journal of Marketing*, 69 (1), 95-113.

Homburg, Christian, Mathias Droll, and Dirk Totzek (2008), "Customer Prioritization: Does It Pay Off, and How Should It Be Implemented?," *Journal of Marketing*, 72 (5), 110-30.

Homburg, Christian, Michael Müller, and Martin Klarmann (2011), "When Should the Customer Really Be King? On the Optimum Level of Salesperson Customer Orientation in Sales Encounters," *Journal of Marketing*, 75 (2), 55-74.

Homburg, Christian, Marcel Stierl, and Torsten Bornemann (2013), "Corporate Social Responsibility in Business-to-Business Markets: How Organizational Customers Account for Supplier Corporate Social Responsibility Engagement," *Journal of Marketing*, 77 (6), 54-72.

Hu, Li-tze and Peter M. Bentler (1999), "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives," *Structural Equation Modeling: A Multidisciplinary Journal*, 6 (1), 1-55.

Hubbard, Raymond (2019), "Will the ASA's Efforts to Improve Statistical Practice Be Successful? Some Evidence to the Contrary," *The American Statistician*, 73 (S1), 31-35.

Huber, Joel and Tülin Erdem (2014), "*JMR* in Transition: Reflections on the 2006–2012 Period," *Journal of Marketing Research*, 51 (1), 133-35.

Hulland, John, Hans Baumgartner, and Keith Marion Smith (2018), "Marketing Survey Research Best Practices: Evidence and Recommendations from a Review of *JAMS* Articles," *Journal of the Academy of Marketing Science*, 46 (1), 92-108.

Hulland, John, Yiu Ho Chow, and Shunyin Lam (1996), "Use of Causal Models in Marketing Research: A Review," *International Journal of Research in Marketing*, 13 (2), 181-97.

Huyghe, Elke, Julie Verstraeten, Maggie Geuens, and Anneleen Van Kerckhove (2017), "Clicks as a Healthy Alternative to Bricks: How Online Grocery Shopping Reduces Vice Purchases," *Journal of Marketing Research*, 54 (1), 61-74.

Inc. (2010), "How to Get Customer Referrals," (accessed March 30, 2020), [available at https://www.inc.com/guides/2010/08/how-to-get-customer-referrals.html].

Isserlis, L. (1917), "The Variation of the Multiple Correlation Coefficient in Samples Drawn from an Infinite Population with Normal Distribution," *Philosophical Magazine Series 6*, 34 (201), 205-20.

Iyengar, Raghuram, Christophe Van den Bulte, and Thomas W. Valente (2011), "Opinion Leadership and Social Contagion in New Product Diffusion," *Marketing Science*, 30 (2), 195-212.

Jaccard, James and Choi K. Wan (1995), "Measurement Error in the Analysis of Interaction Effects between Continuous Predictors Using Multiple Regression: Multiple Indicator and Structural Equation Approaches," *Psychological Bulletin*, 117 (2), 348-57.

Jin, Liyin and Yunhui Huang (2014), "When Giving Money Does Not Work: The Differential Effects of Monetary Versus in-Kind Rewards in Referral Reward Programs," *International Journal of Research in Marketing*, 31 (1), 107-16.

Jöreskog, Karl G. (1971), "Statistical Analysis of Sets of Congeneric Tests," *Psychometrika*, 36 (2), 109-33.

Jöreskog, Karl G. and Dag Sörbom (1989), LISREL VII User's Reference Guide. Mooresville, IN: Scientific Software.

Jöreskog, Karl G. and Fan Yang (1996), "Nonlinear Structural Equation Models: The Kenny-Judd Model with Interaction Effects," in *Advanced Structural Equation Modeling: Issues and Techniques*, G. Marcoulides and R. Schumacker, eds. New York, NY: Psychology Press.

Kahn, Barbara E., Mary Frances Luce, and Stephen M. Nowlis (2006), "Debiasing Insights from Process Tests," *Journal of Consumer Research*, 33 (1), 131-38.

Kalnins, Arturs (2018), "Multicollinearity: How Common Factors Cause Type 1 Errors in Multivariate Regression," *Strategic Management Journal*, 39 (8), 2362-85.

Kelava, Augustin and Benjamin Nagengast (2012), "A Bayesian Model for the Estimation of Latent Interaction and Quadratic Effects When Latent Variables Are Non-Normally Distributed," *Multivariate Behavioral Research*, 47 (5), 717-42.

Kelava, Augustin, Christina S. Werner, Karin Schermelleh-Engel, Helfried Moosbrugger, Dieter Zapf, Yue Ma, Heining Cham, Leona S. Aiken, and Stephen G. West (2011), "Advanced Nonlinear Latent Variable Modeling: Distribution Analytic LMS and QML Estimators of Interaction and Quadratic Effects," *Structural Equation Modeling: A Multidisciplinary Journal*, 18 (3), 465-91.

Kelley, Truman L. (1932), Statistical Method. New York: The Macmillan Company.

Kenny, David A. and Charles M. Judd (1984), "Estimating the Nonlinear and Interactive Effects of Latent Variables," *Psychological Bulletin*, 96 (1), 201-10.

Kerlinger, Fred N. (1977), "The Influence of Research on Education Practice," *Educational Researcher*, 6 (8), 5-12.

Klein, Andreas and Helfried Moosbrugger (2000), "Maximum Likelihood Estimation of Latent Interaction Effects with the LMS Method," *Psychometrika*, 65 (4), 457-74.

Korschun, Daniel, C. B. Bhattacharya, and Scott D. Swain (2014), "Corporate Social Responsibility, Customer Orientation, and the Job Performance of Frontline Employees," *Journal of Marketing*, 78 (3), 20-37.

Kumar, V., J. Andrew Petersen, and Robert P. Leone (2010), "Driving Profitability by Encouraging Customer Referrals: Who, When, and How," *Journal of Marketing*, 74 (5), 1-17.

Lastovicka, John L. and Kanchana Thamodaran (1991), "Common Factor Score Estimates in Multiple Regression Problems," *Journal of Marketing Research*, 28 (1), 105-12.

Le, Huy, Frank L. Schmidt, James K. Harter, and Kristy J. Lauver (2010), "The Problem of Empirical Redundancy of Constructs in Organizational Research: An Empirical Investigation," *Organizational Behavior and Human Decision Processes*, 112 (2), 112-25.

Lee, Clarence, Elie Ofek, and Thomas J. Steenburgh (2018), "Personal and Social Usage: The Origins of Active Customers and Ways to Keep Them Engaged," *Management Science*, 64 (6), 2473-95.

Leskovec, Jure, Lada A. Adamic, and Bernardo A. Huberman (2007), "The Dynamics of Viral Marketing," *ACM Transactions on the Web*, 1 (1), 1-39.

Lin, Guan-Chyun, Zhonglin Wen, Herbert W. Marsh, and Huey-Shyan Lin (2010), "Structural Equation Models of Latent Interactions: Clarification of Orthogonalizing and Double-Mean-Centering Strategies," *Structural Equation Modeling: A Multidisciplinary Journal*, 17 (3), 374-91.

Little, Todd D., James A. Bovaird, and Keith F. Widaman (2006), "On the Merits of Orthogonalizing Powered and Product Terms: Implications for Modeling Interactions among Latent Variables," *Structural Equation Modeling: A Multidisciplinary Journal*, 13 (4), 497-519.

Little, Todd D., William A. Cunningham, Golan Shahar, and Keith F. Widaman (2002), "To Parcel or Not to Parcel: Exploring the Question, Weighing the Merits," *Structural Equation Modeling: A Multidisciplinary Journal*, 9 (2), 151-73.

Lomax, Richard G. (1986), "The Effect of Measurement Error in Structural Equation Modeling," *The Journal of Experimental Education*, 54 (3), 157-62.

Lu, Irene R. R., Ernest Kwan, D. Roland Thomas, and Marzena Cedzynski (2011), "Two New Methods for Estimating Structural Equation Models: An Illustration and a Comparison with Two Established Methods," *International Journal of Research in Marketing*, 28 (3), 258-68.

Lusch, Robert F. and James R. Brown (1996), "Interdependency, Contracting, and Relational Behavior in Marketing Channels," *Journal of Marketing*, 60 (4), 19-38.

Ma, Jingjing and Neal J. Roese (2014), "The Maximizing Mind-Set," *Journal of Consumer Research*, 41 (1), 71-92.

MacKenzie, Scott B. (2001), "Opportunities for Improving Consumer Research through Latent Variable Structural Equation Modeling," *Journal of Consumer Research*, 28 (1), 159-66.

MacKenzie, Scott B. and Philip M. Podsakoff (2012), "Common Method Bias in Marketing: Causes, Mechanisms, and Procedural Remedies," *Journal of Retailing*, 88 (4), 542-55.

MacKenzie, Scott B., Philip M. Podsakoff, and Nathan P. Podsakoff (2011), "Construct Measurement and Validation Procedures in *MIS* and Behavioral Research: Integrating New and Existing Techniques," *MIS Quarterly*, 35 (2), 293-334.

Maclnnis, Deborah J., Christine Moorman, and Bernard J. Jaworski (1991), "Enhancing and Measuring Consumers' Motivation, Opportunity, and Ability to Process Brand Information from Ads," *Journal of Marketing*, 55 (4), 32-53.

Marsh, Herbert W., Zhonglin Wen, and Kit-Tai Hau (2004), "Structural Equation Models of Latent Interactions: Evaluation of Alternative Estimation Strategies and Indicator Construction," *Psychological Methods*, 9 (3), 275-300.

Martin, Kelly D., Abhishek Borah, and Robert W. Palmatier (2017), "Data Privacy: Effects on Customer and Firm Performance," *Journal of Marketing*, 81 (1), 36-58.

Martínez-López, Francisco J., Juan C. Gázquez-Abad, and Carlos M. P. Sousa (2013), "Structural Equation Modelling in Marketing and Business Research," *European Journal of Marketing*, 47 (1/2), 115-52.

Mason, Charlotte H. and William D. Perreault (1991), "Collinearity, Power, and Interpretation of Multiple Regression Analysis," *Journal of Marketing Research*, 28 (3), 268-80.

Maxham III, James G. and Richard G. Netemeyer (2002), "A Longitudinal Study of Complaining Customers' Evaluations of Multiple Service Failures and Recovery Efforts," *Journal of Marketing*, 66 (4), 57-71.

Maxwell, Scott E. and Harold D. Delaney (1993), "Bivariate Median Splits and Spurious Statistical Significance," *Psychological Bulletin*, 113 (1), 181-90.

McDonald, R. (1999), "Test Theory: A Unified Treatment." New York: Psychology Press.

McDonald, Roderick P. and E. J. Burr (1967), "A Comparison of Four Methods of Constructing Factor Scores," *Psychometrika*, 32 (4), 381-401.

McFerran, Brent and Anirban Mukhopadhyay (2013), "Lay Theories of Obesity Predict Actual Body Mass," *Psychological Science*, 24 (8), 1428-36.

Mende, Martin, Ruth N. Bolton, and Mary Jo Bitner (2013), "Decoding Customer–Firm Relationships: How Attachment Styles Help Explain Customers' Preferences for Closeness, Repurchase Intentions, and Changes in Relationship Breadth," *Journal of Marketing Research*, 50 (1), 125-42.

Moorman, Christine, Harald J. van Heerde, C. Page Moreau, and Robert W. Palmatier (2019), "*JM* as a Marketplace of Ideas," *Journal of Marketing*, 83 (1), 1-7.

Moosbrugger, Helfried, Karin Schermelleh-Engel, Augustin Kelava, and Andreas G. Klein (2009), "Testing Multiple Nonlinear Effects in Structural Equation Modeling: A Comparison of Alternative Estimation Approaches," in *Structural Equation Modeling in Educational*

Research: Concepts and Applications, Timothy Teo and Myint Swe Khine, eds. Rotterdam, NL: Sense Publishers.

Moosbrugger, Helfried, Karin Schermelleh-Engel, and Andreas Klein (1997), "Methodological Problems of Estimating Latent Interaction Effects," *Methods of Psychological Research Online*, 2 (2), 95-111.

Müller-Stewens, Jessica, Tobias Schlager, Gerald Häubl, and Andreas Herrmann (2017), "Gamified Information Presentation and Consumer Adoption of Product Innovations," *Journal of Marketing*, 81 (2), 8-24.

Muthén, Bengt O., Linda K. Muthén, and Tihomir Asparouhov (2016), Regression and Mediation Analysis Using Mplus (1st ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, Linda K. and Bengt O. Muthén (2002), "How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power," *Structural Equation Modeling: A Multidisciplinary Journal*, 9 (4), 599-620.

---- (2018), Mplus User's Guide (8th ed.). Los Angeles, CA: Muthén & Muthén.

Narver, John C. and Stanley F. Slater (1990), "The Effect of a Market Orientation on Business Profitability," *Journal of Marketing*, 54 (4), 20-35.

Ng, Jacky C. K. and Wai Chan (2020), "Latent Moderation Analysis: A Factor Score Approach," *Structural Equation Modeling: A Multidisciplinary Journal*, 27 (4), 629-48.

Nickerson, R. S. (2000), "Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy," *Psychological Methods*, 5 (2), 241-301.

Niles, Henry E. (1922), "Correlation, Causation and Wright's Theory of 'Path Coefficients'," *Genetics*, 7 (3), 258-73.

Paley, Anna, Stephanie M. Tully, and Eesha Sharma (2018), "Too Constrained to Converse: The Effect of Financial Constraints on Word of Mouth," *Journal of Consumer Research*, 45 (5), 889-905.

Park, C. Whan, Deborah J. Macinnis, Joseph Priester, Andreas B. Eisingerich, and Dawn Iacobucci (2010), "Brand Attachment and Brand Attitude Strength: Conceptual and Empirical Differentiation of Two Critical Brand Equity Drivers," *Journal of Marketing*, 74 (6), 1-17.

Peighambari, Kaveh, Setayesh Sattari, Arash Kordestani, and Pejvak Oghazi (2016), "Consumer Behavior Research: A Synthesis of the Recent Literature," *SAGE Open*, 6 (2), 1-9.

Perezgonzalez, Jose D. (2015), "Fisher, Neyman-Pearson or NHST? A Tutorial for Teaching Data Testing," *Frontiers in Psychology*, 6 (223), 1-11.

Peter, J. Paul (1981), "Construct Validity: A Review of Basic Issues and Marketing Practices," *Journal of Marketing Research*, 18 (2), 133-45.

Peterson, Robert A. (1994), "A Meta-Analysis of Cronbach's Coefficient Alpha," *Journal of Consumer Research*, 21 (2), 381-91.

Peterson, Robert A. and William R. Wilson (1992), "Measuring Customer Satisfaction: Fact and Artifact," *Journal of the Academy of Marketing Science*, 20 (1), 61.

Petrescu, Maria (2013), "Marketing Research Using Single-Item Indicators in Structural Equation Models," *Journal of Marketing Analytics*, 1 (2), 99-117.

Pieters, Rik (2017), "Meaningful Mediation Analysis: Plausible Causal Inference and Informative Communication," *Journal of Consumer Research*, 44 (3), 692-716.

Ping, Robert A. (1995), "A Parsimonious Estimating Technique for Interaction and Quadratic Latent Variables," *Journal of Marketing Research*, 32 (3), 336-47.

Podsakoff, Philip M., Scott B. MacKenzie, and Nathan P. Podsakoff (2016), "Recommendations for Creating Better Concept Definitions in the Organizational, Behavioral, and Social Sciences," *Organizational Research Methods*, 19 (2), 159-203.

Provine, William B (1992), "The R. A. Fisher-Sewall Wright Controversy," in *The Founders of Evolutionary Genetics: A Centenary Reappraisal*, Sahotra Sarkar, ed. Dordrecht, The Netherlands: Springer.

Provine, William B. (1989), Sewall Wright and Evolutionary Biology. Chicago: University of Chicago Press.

Quine, W. V. and J. S. Ullian (1978), The Web of Belief (2nd ed.). New York: Random House.

R Core Team (2019), "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing.

Ramaseshan, B., Jochen Wirtz, and Dominik Georgi (2017), "The Enhanced Loyalty Drivers of Customers Acquired through Referral Reward Programs," *Journal of Service Management*, 28 (4), 687-706.

Reichardt, Charles S. and S. C. Coleman (1995), "The Criteria for Convergent and Discriminant Validity in a Multitrait-Multimethod Matrix," *Multivariate Behavioral Research*, 30 (4), 513-38.

Richins, Marsha L. and Scott Dawson (1992), "A Consumer Values Orientation for Materialism and Its Measurement: Scale Development and Validation," *Journal of Consumer Research*, 19 (3), 303-16.

Ringle, Christian M., Marko Sarstedt, and Detmar W. Straub (2012), "Editor's Comments: A Critical Look at the Use of PLS-SEM in *MIS Quarterly*," *MIS Quarterly*, 36 (1), iii-xiv.

Roberts, Seth and Harold Pashler (2000), "How Persuasive Is a Good Fit? A Comment on Theory Testing," *Psychological Review*, 107 (2), 358-67.

Rosenthal, R. and M. R. DiMatteo (2001), "Meta-Analysis: Recent Developments in Quantitative Methods for Literature Reviews," *Annual Review of Psychology*, 52 (1), 59-82.

Rosseel, Yves (2012), "lavaan: An R Package for Structural Equation Modeling," *Journal of Statistical Software*, 48 (2), 1-36.

Rossi, Joseph S. (1990), "Statistical Power of Psychological Research: What Have We Gained in 20 Years?," *Journal of Consulting Clinical Psychology*, 58 (5), 646-56.

Rossiter, John R. (2002), "The C-OAR-SE Procedure for Scale Development in Marketing," *International Journal of Research in Marketing*, 19 (4), 305-35.

Russell-Bennett, Rebekah and Steve Baron (2015), "Publishing in *JSM* Part 1: Making a Contribution," *Journal of Services Marketing*, 29 (3), 1-4.

Ryu, Gangseog and Lawrence Feick (2007), "A Penny for Your Thoughts: Referral Reward Programs and Referral Likelihood," *Journal of Marketing*, 71 (1), 84-94.

Saunders, David R. (1955), "The 'Moderator Variable' as a Useful Tool in Prediction.," in Proceedings of the 1954 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service.

Schermelleh-Engel, Karin, Andreas Klein, and Helfried Moosbrugger (1998), "Estimating Nonlinear Effects Using a Latent Moderated Structural Equations Approach.," in *Interaction and Nonlinear Effects in Structural Equation Modeling*, Randall E. Schumacker and George A. Marcoulides, eds. Mahwah, NJ: Lawrence Erlbaum Associates.

Schmitt, Philipp, Bernd Skiera, and Christophe Van den Bulte (2011), "Referral Programs and Customer Value," *Journal of Marketing*, 75 (1), 46-59.

Schoemann, Alexander M., Patrick Miller, Sunthud Pornprasertmanit, and Wei Wu (2014), "Using Monte Carlo Simulations to Determine Power and Sample Size for Planned Missing Designs," *International Journal of Behavioral Development*, 38 (5), 471-79.

Schroll, Roland, Benedikt Schnurr, and Dhruv Grewal (2018), "Humanizing Products with Handwritten Typefaces," *Journal of Consumer Research*, 45 (3), 648-72.

Seiders, Kathleen, Glenn B. Voss, Dhruv Grewal, and Andrea L. Godfrey (2005), "Do Satisfied Customers Buy More? Examining Moderating Influences in a Retailing Context," *Journal of Marketing*, 69 (4), 26-43.

Shaffer, Jonathan A., David DeGeest, and Andrew Li (2016), "Tackling the Problem of Construct Proliferation: A Guide to Assessing the Discriminant Validity of Conceptually Related Constructs," *Organizational Research Methods*, 19 (1), 80-110.

Shah, Rachna and Susan Meyer Goldstein (2006), "Use of Structural Equation Modeling in Operations Management Research: Looking Back and Forward," *Journal of Operations Management*, 24 (2), 148-69.

Shen, Hao and Jaideep Sengupta (2018), "Word of Mouth Versus Word of Mouse: Speaking About a Brand Connects You to It More Than Writing Does," *Journal of Consumer Research*, 45 (3), 595-614.

Sheth, Jagdish N. (1971), "Word-of-Mouth in Low-Risk Innovations," *Journal of Advertising Research*, 11 (3), 15-18.

Shook, Christopher L., David J. Ketchen, G. Tomas M. Hult, and K. Michele Kacmar (2004), "An Assessment of the Use of Structural Equation Modeling in Strategic Management Research," *Strategic Management Journal*, 25 (4), 397-404.

Siemsen, Enno, Aleda Roth, and Pedro Oliveira (2010), "Common Method Bias in Regression Models with Linear, Quadratic, and Interaction Effects," *Organizational Research Methods*, 13 (3), 456-76.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn (2011), "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, 22 (11), 1359-66.

Skrondal, Anders and Petter Laake (2001), "Regression among Factor Scores," *Psychometrika*, 66 (4), 563-75.

Smith, Gregory T. (2005), "On Construct Validity: Issues of Method and Measurement," *Psychological Assessment*, 17 (4), 396.

Spearman, C. (1904), "The Proof and Measurement of Association between Two Things," *The American Journal of Psychology*, 15 (1), 72-101.

Spencer, Steven J., Mark P. Zanna, and Geoffrey T. Fong (2005), "Establishing a Causal Chain: Why Experiments Are Often More Effective Than Mediational Analyses in Examining Psychological Processes," *Journal of Personality and Social Psychology*, 89 (6), 845-51.

Steenkamp, Jan-Benedict E. M. and Hans C. M. van Trijp (1991), "The Use of LISREL in Validating Marketing Constructs," *International Journal of Research in Marketing*, 8 (4), 283-99.

Steffel, Mary and Elanor F. Williams (2018), "Delegating Decisions: Recruiting Others to Make Choices We Might Regret," *Journal of Consumer Research*, 44 (5), 1015-32.

Strauss, Milton E. and Gregory T. Smith (2009), "Construct Validity: Advances in Theory and Methodology," *Annual Review of Clinical Psychology*, 5, 1-25.

Sundaram, Dinesh S., Kaushik Mitra, and Cynthia Webster (1998), "Word-of-Mouth Communications: A Motivational Analysis," *Advances in Consumer Research*, 25, 527-31.

Szucs, Denes and John P. A. Ioannidis (2017a), "When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment," *Frontiers in Human Neuroscience*, 11 (390), 1-21.

Szucs, Denes and John P.A. Ioannidis (2017b), "Empirical Assessment of Published Effect Sizes and Power in the Recent Cognitive Neuroscience and Psychology Literature," *PloS Biology*, 15 (3), 1-18.

Tellis, Gerard J. (2017), "Interesting and Impactful Research: On Phenomena, Theory, and Writing," *Journal of the Academy of Marketing Science*, 45 (1), 1-6.

Tepper, Bennett J. and Kelly Tepper (1993), "The Effects of Method Variance within Measures," *The Journal of Psychology*, 127 (3), 293-302.

Tesla (2019), "Tesla's New Customer Referral Program," (accessed March 5, 2020), [available at https://www.tesla.com/blog/teslas-new-customer-referral-program].

Tian, Kelly Tepper, William O. Bearden, and Gary L. Hunter (2001), "Consumers' Need for Uniqueness: Scale Development and Validation," *Journal of Consumer Research*, 28 (1), 50-66.

Tofighi, Davood and David P. MacKinnon (2016), "Monte Carlo Confidence Intervals for Complex Functions of Indirect Effects," *Structural Equation Modeling: A Multidisciplinary Journal*, 23 (2), 194-205.

Trusov, Michael, Randolph E. Bucklin, and Koen Pauwels (2009), "Effects of Word-of-Mouth Versus Traditional Marketing: Findings from an Internet Social Networking Site," *Journal of Marketing*, 73 (5), 90-102.

Tucker, Ledyard R. (1971), "Relations of Factor Score Estimates to Their Use," *Psychometrika*, 36 (4), 427-36.

Umbach, Nora, Katharina Naumann, Holger Brandt, and Augustin Kelava (2017), "Fitting Nonlinear Structural Equation Models in R with Package nlsem," *Journal of Statistical Software*, 77 (1), 1-20.

Uncles, Mark D., Robert East, and Wendy Lomax (2013), "Good Customers: The Value of Customers by Mode of Acquisition," *Australasian Marketing Journal*, 21 (2), 119-25.

Vale, C. David and Vincent A. Maurelli (1983), "Simulating Multivariate Nonnormal Distributions," *Psychometrika*, 48 (3), 465-71.

Van den Bulte, Christophe, Emanuel Bayer, Bernd Skiera, and Philipp Schmitt (2018), "How Customer Referral Programs Turn Social Capital into Economic Capital," *Journal of Marketing Research*, 55 (1), 132-46.

Van Laer, Tom, Jennifer Edson Escalas, Stephan Ludwig, and Ellis A. Van den Hende (2018), "What Happens in Vegas Stays on Tripadvisor? A Theory and Technique to Understand Narrativity in Consumer Reviews," *Journal of Consumer Research*, 46 (2), 267-85.

Varadarajan, P. Rajan (1996), "From the Editor: *The Journal of Marketing*, 1993 to 1996," *Journal of Marketing*, 60 (4), 1-2.

Venn, John (1880), "On the Diagrammatic and Mechanical Representation of Propositions and Reasonings," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 10 (59), 1-18.

Verhoef, Peter C. and Bas Donkers (2005), "The Effect of Acquisition Channels on Customer Loyalty and Cross-Buying," *Journal of Interactive Marketing*, 19 (2), 31-43.

Verlegh, Peeter W. J., Gangseog Ryu, Mirjam A. Tuk, and Lawrence Feick (2013), "Receiver Responses to Rewarded Referrals: The Motive Inferences Framework," *Journal of the Academy of Marketing Science*, 41 (6), 669-82.

Villanueva, Julian, Shijin Yoo, and Dominique M. Hanssens (2008), "The Impact of Marketing-Induced Versus Word-of-Mouth Customer Acquisition on Customer Equity Growth," *Journal of Marketing Research*, 45 (1), 48-59.

Viswanathan, Vijay, Sebastian Tillmanns, Manfred Krafft, and Daniel Asselmann (2018), "Understanding the Quality–Quantity Conundrum of Customer Referral Programs: Effects of Contribution Margin, Extraversion, and Opinion Leadership," *Journal of the Academy of Marketing Science*, 46 (6), 1108-32.

Völckner, Franziska and Henrik Sattler (2006), "Drivers of Brand Extension Success," *Journal of Marketing*, 70 (2), 18-34.

Von Wangenheim, Florian and Tomás Bayón (2004), "Satisfaction, Loyalty and Word of Mouth within the Customer Base of a Utility Provider: Differences between Stayers, Switchers and Referral Switchers," *Journal of Consumer Behaviour*, 3 (3), 211-20.

Voorhees, Clay M., Michael K. Brady, Roger Calantone, and Edward Ramirez (2016), "Discriminant Validity Testing in Marketing: An Analysis, Causes for Concern, and Proposed Remedies," *Journal of the Academy of Marketing Science*, 44 (1), 119-34.

Voss, Glenn B. and Zannie Giraud Voss (2000), "Strategic Orientation and Firm Performance in an Artistic Environment," *Journal of Marketing*, 64 (1), 67-83.

Wang, Wenbo, Aradhna Krishna, and Brent McFerran (2017), "Turning Off the Lights: Consumers' Environmental Efforts Depend on Visible Efforts of Firms," *Journal of Marketing Research*, 54 (3), 478-94.

Wanous, John P. and Michael J. Hudy (2001), "Single-Item Reliability: A Replication and Extension," *Organizational Research Methods*, 4 (4), 361-75.

Westbrook, Robert A. (1980), "A Rating Scale for Measuring Product/Service Satisfaction," *Journal of Marketing*, 44 (4), 68-72.

Westfall, Jacob and Tal Yarkoni (2016), "Statistically Controlling for Confounding Constructs Is Harder Than You Think," *PLoS ONE*, 11 (3), 1-22.

Wetzer, Inge M., Marcel. Zeelenberg, and Rik Pieters (2007), "'Never Eat in That Restaurant, I Did!': Exploring Why People Engage in Negative Word-of-Mouth Communication," *Psychology & Marketing*, 24 (8), 661-80.

Whetten, David A. (1989), "What Constitutes a Theoretical Contribution?," *Academy of Management Review*, 14 (4), 490-95.

Wilcox, Keith, Anne L. Roggeveen, and Dhruv Grewal (2011), "Shall I Tell You Now or Later? Assimilation and Contrast in the Evaluation of Experiential Products," *Journal of Consumer Research*, 38 (4), 763-73.

Wolfle, Lee M. (1999), "Sewall Wright on the Method of Path Coefficients: An Annotated Bibliography," *Structural Equation Modeling: A Multidisciplinary Journal*, 6 (3), 280-91.

Wright, Sewall (1921), "Correlation and Causation," *Journal of Agricultural Research*, 20, 557-85.

---- (1934), "The Method of Path Coefficients," *The Annals of Mathematical Statistics*, 5 (3), 161-215.

---- (1918), "On the Nature of Size Factors," *Genetics*, 3 (4), 367-74.

---- (1920), "The Relative Importance of Heredity and Environment in Determining the Piebald Pattern of Guinea-Pigs," *Proceedings of the National Academy of Sciences of the United States of America*, 6 (6), 320-32.

Yuan, Ke-Hai, Ying Cheng, and Wei Zhang (2010), "Determinants of Standard Errors of MLEs in Confirmatory Factor Analysis," *Psychometrika*, 75 (4), 633-48.

Zhang, Jie, Michel Wedel, and Rik Pieters (2009), "Sales Effects of Attention to Feature Advertisements: A Bayesian Mediation Analysis," *Journal of Marketing Research*, 46 (5), 669-81.

CENTER DISSERTATION SERIES

CentER for Economic Research, Tilburg University, the Netherlands

No.	Author	Title	ISBN	Published
579	Julius Rüschenpöhler	Behavioural Perspectives on Subsistence Entrepreneurship in Emerging Markets	978 90 5668 580 5	January 2019
580	Khulan Altangerel	Essays on Immigration Policy	978 90 5668 581 2	January 2019
581	Kun Zheng	Essays on Duration Analysis and Labour Economics	978 90 5668 582 9	January 2019
582	Tatiana Zabara	Evolution of Entrepreneurial Teams in Technology-Based New Ventures	978 90 5668 583 6	February 2019
583	Yifan Yu	Essays on Mixed Hitting-Time Models	978 90 5668 584 3	April 2019
584	Daniel Martinez Martin	Unpacking Product Modularity, Innovation in Distributed Innovation Teams	978 90 5668 585 0	April 2019
585	Katalin Katona	Managed Competition in Practice Lessons for Healthcare Policy	978 90 5668 586 7	April 2019
586	Serhan Sadikoglu	Essays in Econometric Theory	978 90 5668 587 4	May 2019
587	Hoang Yen Nguyen	Emotions and Strategic Interactions	978 90 5668 588 1	May 2019
588	Ties de Kok	Essays on reporting and information processing	978 90 5668 589 8	May 2019
589	Yusiyu Wang	Regulation, Protest, and Spatial Economics	978 90 5668 590 4	June 2019
590	Ekaterina Neretina	Essays in Corporate Finance, Political Economy, and Competition	978 90 5668 591 1	June 2019
591	Ruth Wandhöfer	Technology innovation in Financial Markets: Implications for Money, Payments and Settlement Finality	978 90 5668 592 8	June 2019

No.	Author	Title	ISBN	Published
592	Andinet Worku Gebreselassie	On communicating about taboo social issues in least developed countries: The case of Ethiopia	978 90 5668 593 5	June 2019
593	Filip Bekjarovski	Active Investing	978 90 5668 594 2	June 2019
594	Miguel Sarmiento	Essays on Banking, Financial Intermediation and Financial Markets	978 90 5668 595 9	June 2019
595	Xiaoyin Ma	Essays on Alternative Investements	978 90 5668 596 6	June 2019
596	Victor van Pelt	A Dynamic View of Management Accounting Systems	978 90 5668 597 3	June 2019
597	Shuai Chen	Marriage, Minorities, and Mass Movements	978 90 5668 598 0	July 2019
598	Ben Gans	Stabilisation operations as complex systems: order and chaos in the interoperability continuum	978 90 5668 599 7	July 2019
599	Mulu Hundera	Role Conflict, Coping Strategies and Female Entrepreneurial Success in Sub-Saharan Africa	978 90 5668 600 0	August 2019
600	Hao Hu	The Quadratic Shortest Path Problem – Theory and Computations	978 90 5668 601 7	September 2019
601	Emerson Erik Schmitz	Essays on Banking and International Trade	978 90 5668 602 4	September 2019
602	Olga Kuryatnikova	The many faces of positivity to approximate structured optimization problems	978 90 5668 603 1	September 2019
603	Sander Gribling	Applications of optimization to factorization ranks and quantum information theory	978 90 5668 604 8	September 2019
604	Camille Hebert	Essays on Corporate Ownership and Human Capital	978 90 5668 605 5	October 2019
605	Gabor Neszveda	Essays on Behavioral Finance	978 90 5668 606 2	October 2019

No.	Author	Title	ISBN	Published
606	Ad van Geesbergen	Duurzame schaarste - Een kritische analyse van twee economische duurzaamheids-paradigma's geïnspireerd door de filosofie van Dooyeweerd	978 90 5668 607 9	October 2019
607	Richard T. Mason	Digital Enrollment Architecture and Retirement Savings Decisions: Evidence from the Field	978 90 5668 608 6	November 2019
608	Ron Triepels	Anomaly Detection in the Shipping and Banking Industry	978 90 5668 609 3	November 2019
609	Feng Fang	When performance shortfall arises, contract or trust? A multi-method study of the impact of contractual and relation governances on performance in Public-Private Partnerships	978 90 5668 610 9	November 2019
610	Yasir Dewan	Corporate Crime and Punishment: The Role of Status and Ideology	978 90 5668 611 6	November 2019
611	Mart van Hulten	Aiming for Well-Being through Taxation: A Framework of Caution and Restraint for States	978 90 5668 612 3	December 2019
612	Carlos Sandoval Moreno	Three essays on poverty measurement and risk protection	978 90 5668 613 0	December 2019
613	Harmke de Groot	Core strength or Achilles' heel: Organizational competencies and the performance of R&D collaborations	978 90 5668 614 7	December 2019
614	Peter Brok	Essays in Corporate Finance and Corporate Taxation	978 90 5668 615 4	December 2019
615	Pascal Böni	On the Pricing, Wealth Effects and Return of Private Market Debt	978 90 5668 616 1	December 2019
616	Ana Martinovici	Revealing Attention: How Eye Movements Predict Brand Choice and Moment of Choice	978 90 5668 617 8	December 2019
617	Matjaz Maletic	Essays on international finance and empirical asset pricing	978 90 5668 618 5	January 2020
618	Zilong Niu	Essays on Asset Pricing and International Finance	978 90 5668 619 2	January 2020

No.	Author	Title	ISBN	Published
619	Bjorn Lous	On free markets, income inequality, happiness and trust	978 90 5668 620 8	January 2020
620	Clemens Fiedler	Innovation in the Digital Age: Competition, Cooperation, and Standardization	978 90 5668 621 5	June 2020
621	Andreea Popescu	Essays in Asset Pricing and Auctions	978 90 5668 622 2	June 2020
622	Miranda Stienstra	The Determinants and Performance Implications of Alliance Partner Acquisition	978 90 5668 623 9	June 2020
623	Lei Lei	Essays on Labor and Family Economics in China	978 90 5668 624 6	May 2020
624	Farah Arshad	Performance Management Systems in Modern Organizations	978 90 5668 625 3	June 2020
625	Yi Zhang	Topics in Economics of Labor, Health, and Education	978 90 5668 626 0	June 2020
626	Emiel Jerphanion	Essays in Economic and Financial decisions of Households	978 90 5668 627 7	July 2020
627	Richard Heuver	Applications of liquidity risk discovery using financial market infrastructures transaction archives	978 90 5668 628 4	September 2020
628	Mohammad Nasir Nasiri	Essays on the Impact of Different Forms of Collaborative R&D on Innovation and Technological Change	978 90 5668 629 1	August 2020
629	Dorothee Hillrichs	On inequality and international trade	978 90 5668 630 7	September 2020
630	Roland van de Kerkhof	It's about time: Managing implementation dynamics of condition-based maintenance	978 90 5668 631 4	October 2020
631	Constant Pieters	Process Analysis for Marketing Research	978 90 5668 632 1	December 2020

This dissertation consists of three essays on process analysis for marketing research. The first essay (Chapter 2) investigates the referral reinforcement effect: referred customers have a higher inclination of making referrals than non-referred customers have. Four studies quantify the referral reinforcement effect and mediation analyses decompose it in satisfaction-mediated and non-satisfaction-mediated pathways. A final study explores customer lay beliefs about potential drivers of the referral reinforcement effect. Implications for marketing theory and practice are discussed. The second essay (Chapter 3) compares six existing moderation methods in the face of random measurement error. A quantitative literature review documents their use in marketing research and Monte Carlo simulations assess their performance. Recommendations for future usage of the moderation methods are provided. The third essay (Chapter 4) focuses on discriminant validity as a precondition for meaningful process analysis in marketing research. It extends existing bivariate criteria for discriminant validity with multivariate discriminant validity criteria. Case studies taken from a quantitative literature review apply the bivariate and multivariate discriminant validity criteria. An online application is developed to increase the accessibility of the criteria.

Constant Pieters (Oostburg, The Netherlands, 1990) received a BSc. degree in Business Administration (2011), an MSc. degree in Marketing Research (2012), and a Research MSc. degree in Business: Marketing (2014), all at Tilburg University. Before teaching at the Department of Marketing at Tilburg University, he started there as a Ph.D. Candidate in 2014. He is incoming Lecturer at the University of New South Wales, Australia.

ISBN: 978 90 5668 632 1
DOI: 10.26116/center-lis-2009

TILBURG SCHOOL OF ECONOMICS AND MANAGEMENT

CentER Dissertation Series